Seemingly Unrelated Manuscripts: Experiments on Human Behavior

Dissertation
submitted to the Faculty of Economics,
Business Administration and Information Technology
of the University of Zurich

to obtain the degree of
Doctor of Philosophy
in Economics

presented by

Tony Brett Williams
from United States of America

approved in July 2014 at the request of
Prof. Dr. Ernst Fehr
Prof. Dr. Roberto Weber

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 16.07.2014

Chairman of the Doctoral Board: Prof. Dr. Josef Zweimüller

For Donald Clark Hodges and Jane Casey. . . none of this would have been possible without you.

Acknowledgments

I would first like to acknowledge my advisors, Ernst Fehr and Roberto Weber. I also want to acknowledge the support of my closest collaborators: Chris Burke, Bastiaan Oud, and Philippe Tobler. Bastiaan and I worked closely together conducting the experiments and analyzing the data for one of the chapters. Chris and Philippe are responsible for training me how to analyze and interpret functional neuroimaging data; without their guidance and support on the third chapter, I would not have been able to finish this dissertation. Finally, I would like to thank Tim Salmon for his unending support as a friend and mentor over the last decade, starting when I made the decision to pursue economics.

Contents

Chapter 1

Introduction

> *"[. . .] economics has increasingly (if unknowingly for the most part) moved toward an approach that combines the mathematical advances of the last century with three of the methods of the classical economists. From Adam Smith to John Stuart Mill and Karl Marx (and excepting David Ricardo), the classical economists were* nondisciplinary *(the disciplines had not been invented),* concerned about the empirical details *of the social problems of their day, and* modest about the degree of generality to which their theories aspired*."*
>
> – Samuel Bowles (2009, pg. 15)

The title I have given this dissertation, *Seemingly Unrelated Manuscripts: Experiments on Human Behavior*, may sound glib to the reader. Much like seemingly unrelated regressions, however, the chapters of this dissertation can be viewed separately but maintain some underlying correlation across chapters. Superficially, the main connection between the chapters in this dissertation is the common use of experiments with real monetary incentives. More importantly, but perhaps less obvious, is the emphasis on understanding human behavior without being constrained by rigid disciplinary boundaries. In this sense, I am attempting to follow in the footsteps of classical economists as noted in the quote above by Samuel Bowles but also follow the lead of a recent generation of

economists who have transcended their original disciplinary boundaries – most notably by Bowles himself along with Colin Camerer, Ernst Fehr, Herb Gintis – and their colleagues from other disciplines, particularly Robert Boyd, Joe Henrich, and Pete Richerson.

1.1 General background

The chapters of this dissertation fall under two broad themes. Chapters A and B are about understanding human cooperation. Chapter C is about understanding the neural foundations of decision making, in which we focus specifically on risky decision making. Descriptions of the chapter contents are described in more detail in Section 1.2. Here, we present a general background for the two broad themes.

1.1.1 Human cooperation

Human cooperation has long been a topic of interest in the social and biological sciences, and integrating knowledge across fields is crucial in identifying the foundations of cooperative behavior (Fehr and Gächter, 2002; Fehr and Fischbacher, 2003; Henrich, 2004). Due to the confluence of knowledge from these various disciplines, *Science* regarded the puzzle of human cooperation as one of the most fundamentally important open questions upon which we may achieve a greater understanding in the coming decades (Pennisi, 2005).

The social sciences have generally looked for mechanisms that can be used to establish and maintain cooperation. Repeated interactions and punishment – and their combination – are used to sustain cooperation in both natural environments (Ostrom, 1990; Mathew and Boyd, 2011) and laboratory environments (Ostrom et al., 1992; Fehr and Gächter, 2000). It is not surprising that long-term repeated interactions can sustain cooperation; these interactions have long been used in industrial organization to explain the existence of cartels, as the threat of withdrawing future cooperation can serve as an effective deterrent to defect from the cartel in the present time period (Tirole, 1988). What remains surprising and still lacks explanation, however, is the observation that people will

punish non-cooperators in short-term interactions in which the costs of punishment exceed the gains from future cooperation (Gächter et al., 2008) and will even punish in one-shot interactions with no potential future gains, so-called "altruistic punishment" (Fehr and Gächter, 2002).

While the social sciences may be able to find proximate causes to explain human cooperation, ultimate causes must be explained by evolutionary forces (Fehr and Fischbacher, 2003). Kin selection remains an uncontroversial explanation for some forms of altruism (Hamilton, 1964); however, as Trivers (1971) clearly states, biological models of altruism are intended to redefine interactions in such a way that the altruistic act is no longer a pure form of altruism. Such an approach is similar to economic theories of social preferences which redefine an individual's utility function to include features, such as aversion to inequality, that make the "altruistic" act individually optimal and hence "selfish" (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). Several forms of reciprocity, relying on repeated interactions, remain the primary explanation for cooperation in non-related individuals, providing an evolutionary foundation for observations in the social sciences described above (Trivers, 1971; Nowak, 2006; Axelrod and Hamilton, 1981; Axelrod, 2006). Finally, models of multilevel or group selection remain popular but controversial; such models suggest that cooperative groups can develop resources to support increasing population faster than non-cooperative groups, and group-level selective pressure supports cooperation in the long-run (Henrich, 2004; Nowak, 2006).

The approaches discussed so far assume, implicitly or explicitly, that interactions are random. However, in many situations, people have some ability to choose their partners for interaction or to avoid interacting with particular people. Humans are well known to stereotype and discriminate, and many of the social sciences – particulary psychology and sociology – have devoted substantial attention to the topic of discrimination. Assortment, or non-random interaction, can theoretically provide a biological basis for cooperation (Bergstrom, 2003; Eshel and Cavalli-Sforza, 1982). The existence of observable features that could serve as the basis for such assortment, however, are generally considered

implausible (Grafen, 1990; Gardner and West, 2010). Nonetheless, some such features have been identified in non-human species (Gardner and West, 2010), and Stirrat and Perrett (2010) claim to identify a candidate feature in humans faces, the bizygomatic width-to-height ratio. Assortment as an explanation for human cooperation thus remains an open question, but one that is currently being explored.

1.1.2 Neural foundations of value-based decision making

The other broad theme of this dissertation is an attempt to understand the neural basis for decision making. This field, neuroeconomics, remains the topic of much debate in economics (Camerer et al., 2005; Caplin and Schotter, 2008). I avoid addressing the debate here and instead focus on the insights gained from the field and how neuroeconomics may be scientifically important even if it irrelevant for academic economics research. Much like our understanding of human cooperation benefitting from the confluence of ideas from several disciplines, neuroeconomics has progressed quickly due to interaction between neuroscientists, psychologists, and economists (Rangel et al., 2008; Fehr and Rangel, 2011; Camerer, 2013).

Arguably, the prefrontal cortex have received the bulk of attention in the field. One possible reason for this attention is that the prefrontal cortex is a fairly recent evolutionary development and seems involved in many advanced behaviors of both non-human and especially human primates (Teffer and Semendeferi, 2012; Adams et al., 2012). Our current understanding of the brain circuitry of valuation suggests that primary and secondary rewards have distinct but overlapping networks which converge in the ventromedial prefrontal cortex, where they are converted into a common neural currency (Levy and Glimcher, 2011, 2012). The dorsolateral prefrontal cortex (DLPFC), while not completely understood, plays a key role several types of decisions. The DLPFC has long been thought to govern difficult cognitive processes, such as problem solving, but is also crucial in exercising self control and demonstrating socially acceptable behavior (Knoch et al., 2006; Sanfey et al., 2006; Hare et al., 2009). Nonetheless, regions outside of the

prefrontal cortex – such as the temporoparietal junction – are important in value-based decisions in social environments (Morishima et al., 2012).

One of the aims of neuroeconomics is understanding the neural processes in abnormal decision making that characterizes various psychiatric disorders (Montague et al., 2012; Lee, 2013). Mental illness has serious economic consequences. One estimate from the World Economic Forum places the worldwide cost of mental illness at US$2.5 trillion in 2010, two-thirds of which are indirect costs such as lost productivity and income due to worker absenteeism (Bloom et al., 2011). Identifying the neural circuity of value-based decision making in healthy subjects provides a foundation for identifying aberrations present in psychiatric disorders. Currently, we have only a partial understanding of how neuromodulators such as serotonin, norepiniphrine, dopamine, and oxytocin affect behavior (Pitman et al., 1993; Kosfeld et al., 2005; Doya, 2008; Cools, 2012; Crockett and Fehr, 2013). Importantly, serotonin, norepinephrine, and dopamine disturbances have all been implicated in unipolar depression, which is arguably the most common mental illness and has the highest cost (Asberg et al., 1976; Owens and Nemeroff, 1994; Nutt et al., 2006; Dunlop and Nemeroff, 2007; Bloom et al., 2011). Moreover, pharmacological substances affecting these neurotransmitters are the most common treatments for unipolar depression, such as selective serotonin reuptake inhibitors and norepinephrine-dopamine reuptake inhibitors. Thus, neuroeconomics may not only help us understand value-based decision making in normal, healthy individuals but also to identify causes and potential treatments for aberrations in mental illnesses.

1.2 Contents of this dissertation

This section describes the three chapters comprising this dissertation. These chapters are included as appendices to this introductory chapter.

1.2.1 Endogenous emergence of institutions to sustain cooperation

This chapter experimentally examines the roles of "good" people and "good" institutions on cooperation. We first allow a large group (12 subjects) to experience the free-rider problem in a typical public goods game. We then allow subjects to self-select into one of four institutions allowing (or disallowing) different punishment mechanisms and cheap-talk communication. Cooperative people quickly sort into the institutions allowing both punishment – either peer or centralized – and communication. Self-interested people then migrate into these institutions and cooperate by contributing to the public good, minimizing the need for punishment against low contributions. We follow up with a control treatment to disentangle the sorting effects and institutional effects and find support for both effects. The combined effects of sorting into effective institutions quickly and efficiently establishes high rates of cooperation.

1.2.2 What you see is what you get? The effect of facial cues on trust-related behavior

This chapter experimentally examines whether people discriminate in their trust-related behavior and, if so, whether this discrimination is financially beneficial. Oosterhof and Todorov (2008) identify facial features that affect non-incentivized ratings of trustworthiness. We use photographs of real subjects in a modified trust game, along with morphing software, to test whether these perceptions affect behavior in the presence of monetary incentives. We find evidence that many first-movers and second-movers discriminate based on their perceptions, even though only first-movers have a financial incentive to discriminate (conditional on the perceptions being accurate). Furthermore, we find that these perceptions are wildly inaccurate, and average payoffs are no better than the expected payoff from making random choices. Using such facial cues is consistent with an attempt at sustaining cooperation by assortment, but the lack of benefit from such assortment

fails to provide evidence for assortment as an ultimate cause of cooperation; nonetheless, there may be other observable features that are valid cues for assortment.

1.2.3 Neural evidence for computational processes underlying risky decision making

This final chapter attempts to provide some clarity to the use of – and information gained from – various behavioral models of risky decision making to identify neural regions encoding and comparing subjective value. Subjects in our study completed a two-alternative forced choice in which the options were lotteries with monetary payouts. In a first step, we use the three most common models of subjective value in the neuroeconomics literature: expected utility, prospect theory, and a type of mean-variance model which includes skewness. All three models identify the same neural regions, providing robust evidence for the findings of valuation which do not attempt to use alternative models. Portions of the lateral intraparietal area and dorsolateral prefrontal cortex correlate with the sum of behavioral utilities, suggesting that these regions encode subjective value; moreover,the medial orbitofrontal cortex and ventromedial prefrontal cortex correlate with the difference in behavioral utilities, suggesting comparison of subjective value. We then use Bayesian model selection to test whether the best-fitting behavioral model also best explains neural activity in these regions. Using only those subjects whose behavior is most consistent with prospect theory, we find that the neural activity of these subjects is best explained by a mean-variance-skewness model. Finally, we suggest that this apparent discrepancy can potentially be reconciled using theories of optimal foraging and reinforcement learning (both of which are common in the animal learning literature) that can generate reference-dependent and domain-specific risk preferences that are consistent with prospect theory but only require encoding of simple statistical moments such as mean and variance.

1.3 References

(????): .

(????): .

Adams, G. K., K. K. Watson, J. Pearson, and M. L. Platt (2012): "Neuroethology of decision-making," *Current opinion in neurobiology*, 22, 982–989.

Asberg, M., P. Thoren, L. Traskman, L. Bertilsson, and V. Ringberger (1976): "Serotonin depression – a biochemical subgroup within the affective disorders?" *Science*, 191, 478–480.

Axelrod, R. and W. D. Hamilton (1981): "The evolution of cooperation," *Science*, 211, 1390–1396.

Axelrod, R. M. (2006): *The evolution of cooperation*, Basic books.

Bergstrom, T. C. (2003): "The algebra of assortative encounters and the evolution of cooperation," *International Game Theory Review*, 5, 211–228.

Bloom, D., E. Cafiero, E. Jané-Llopis, S. Abrahams-Gessel, L. Bloom, S. Fathima, A. Feigl, T. Gaziano, M. Mowafi, A. Pandya, K. Prettner, L. Rosenberg, B. Seligman, A. Stein, and C. Weinstein (2011): *The Global Economic Burden of Noncommunicable Diseases*, World Economic Forum.

Bolton, G. E. and A. Ockenfels (2000): "ERC: A theory of equity, reciprocity, and competition," *American Economic Review*, 166–193.

Bowles, S. (2009): *Microeconomics: behavior, institutions, and evolution*, Princeton University Press.

Camerer, C. (2013): "Goals, Methods, and Progress in Neuroeconomics," *Annual Review of Econonomics*, 5, 425–455.

Camerer, C., G. Loewenstein, and D. Prelec (2005): "Neuroeconomics: How neuroscience can inform economics," *Journal of Economic Literature*, 9–64.

Caplin, A. and A. Schotter (2008): *The foundations of positive and normative economics: a handbook*, Oxford University Press.

Cools, R. (2012): "Chemical Neuromodulation of Goal-Directed Behavior," in *Cognitive Search: Evolution, Algorithms, and the Brain*, ed. by P. Todd, T. Hills, and T. Robbins, MIT Press.

Crockett, M. J. and E. Fehr (2013): "Pharmacology of Economic and Social Decision-Making," in *Neuroeconomics: Decision-Making and the Brain*, ed. by P. W. Glimcher and E. Fehr, Academic Press.

Doya, K. (2008): "Modulators of decision making," *Nature neuroscience*, 11, 410–416.

Dunlop, B. W. and C. B. Nemeroff (2007): "The role of dopamine in the pathophysiology of depression," *Archives of General Psychiatry*, 64, 327–337.

Eshel, I. and L. L. Cavalli-Sforza (1982): "Assortment of encounters and evolution of cooperativeness," *Proceedings of the National Academy of Sciences*, 79, 1331.

Fehr, E. and U. Fischbacher (2003): "The nature of human altruism," *Nature*, 425, 785–791.

——— (2005): "Altruists with green beards," *Analyse & Kritik*, 27, 73–84.

Fehr, E. and S. Gächter (2000): "Cooperation and punishment in public goods experiments," *American Economic Review*, 90, 980–994.

——— (2002): "Altruistic punishment in humans," *Nature*, 415, 137–140.

Fehr, E. and A. Rangel (2011): "Neuroeconomic foundations of economic choicerecent advances," *Journal of Economic Perspectives*, 25, 3–30.

Fehr, E. and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 114, 817–868.

Frank, R. H. (1987): "If homo economicus could choose his own utility function, would he want one with a conscience?" *American Economic Review*, 593–604.

Gächter, S., E. Renner, and M. Sefton (2008): "The long-run benefits of punishment," *Science*, 322, 1510–1510.

Gardner, A. and S. A. West (2010): "Greenbeards," *Evolution*, 64, 25–38.

Gintis, H., E. A. Smith, and S. Bowles (2001): "Costly signaling and cooperation," *Journal of Theoretical Biology*, 213, 103–119.

Grafen, A. (1990): "Biological signals as handicaps," *Journal of Theoretical Biology*, 144, 517–546.

Hamilton, W. D. (1964): "The genetical evolution of social behaviour. II," *Journal of Theoretical Biology*, 7, 17–52.

Hare, T. A., C. F. Camerer, and A. Rangel (2009): "Self-control in decision-making involves modulation of the vmPFC valuation system," *Science*, 324, 646–648.

Henrich, J. (2004): "Cultural group selection, coevolutionary processes and large-scale cooperation," *Journal of Economic Behavior and Organization*, 53, 3–35.

Knoch, D., A. Pascual-Leone, K. Meyer, V. Treyer, and E. Fehr (2006): "Diminishing reciprocal fairness by disrupting the right prefrontal cortex," *Science*, 314, 829–832.

Kosfeld, M., M. Heinrichs, P. J. Zak, U. Fischbacher, and E. Fehr (2005): "Oxytocin increases trust in humans," *Nature*, 435, 673–676.

Lee, D. (2013): "Decision making: from neuroscience to psychiatry," *Neuron*, 78, 233–248.

Levy, D. J. and P. W. Glimcher (2011): "Comparing apples and oranges: using reward-specific and reward-general subjective value representation in the brain," *Journal of Neuroscience*, 31, 14693–14707.

——— (2012): "The root of all value: a neural common currency for choice," *Current opinion in neurobiology*, 22, 1027–1038.

Mathew, S. and R. Boyd (2011): "Punishment sustains large-scale cooperation in prestate warfare," *Proceedings of the National Academy of Sciences*, 108, 11375–11380.

Montague, P. R., R. J. Dolan, K. J. Friston, and P. Dayan (2012): "Computational psychiatry," *Trends in cognitive sciences*, 16, 72–80.

Morishima, Y., D. Schunk, A. Bruhin, C. C. Ruff, and E. Fehr (2012): "Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism," *Neuron*, 75, 73–79.

Nowak, M. A. (2006): "Five rules for the evolution of cooperation," *Science*, 314, 1560–1563.

Nutt, D. J. et al. (2006): "The role of dopamine and norepinephrine in depression and antidepressant treatment," *Journal of Clinical Psychiatry*, 67, 3.

Oosterhof, N. N. and A. Todorov (2008): "The functional basis of face evaluation," *Proceedings of the National Academy of Sciences*, 105, 11087–11092.

Ostrom, E. (1990): *Governing the commons: The evolution of institutions for collective action*, Cambridge University Press.

Ostrom, E., J. Walker, and R. Gardner (1992): "Covenants with and without a sword: Self-governance is possible," *American Political Science Review*, 404–417.

Owens, M. J. and C. B. Nemeroff (1994): "Role of serotonin in the pathophysiology of depression: focus on the serotonin transporter." *Clinical chemistry*, 40, 288–295.

Pennisi, E. (2005): "How did cooperative behavior evolve?" *Science*, 309, 93–93.

Pitman, R. K., S. P. Orr, and N. B. Lasko (1993): "Effects of intranasal vasopressin and oxytocin on physiologic responding during personal combat imagery in Vietnam veterans with posttraumatic stress disorder," *Psychiatry Research*, 48, 107–117.

Rangel, A., C. Camerer, and P. R. Montague (2008): "A framework for studying the neurobiology of value-based decision making," *Nature Reviews Neuroscience*, 9, 545–556.

Sanfey, A. G., G. Loewenstein, S. M. McClure, and J. D. Cohen (2006): "Neuroeconomics: cross-currents in research on decision-making," *Trends in cognitive sciences*, 10, 108–116.

Stirrat, M. and D. I. Perrett (2010): "Valid Facial Cues to Cooperation and Trust Male Facial Width and Trustworthiness," *Psychological Science*, 21, 349–354.

Teffer, K. and K. Semendeferi (2012): "Human prefrontal cortex: Evolution, development, and pathology," in *Progress in Brain Research*, ed. by M. A. Hofman and D. Falk, Elsevier, vol. 195, 191–218.

Tirole, J. (1988): *The theory of industrial organization*, MIT press.

Trivers, R. L. (1971): "The evolution of reciprocal altruism," *Quarterly Review of Biology*, 46, 35–57.

Appendix A

Endogenous emergence of institutions to sustain cooperation

This chapter is being prepared for submission to a leading economics journal and follows standard formatting for such journals. Work in this chapter was conducted with Ernst Fehr. This chapter was written by Tony Williams and Ernst Fehr.

A.1 Abstract

Formal and informal institutions, such as laws and social norms, are pervasive in daily life. They help maintain cooperation by coordinating and constraining individuals' behaviors. However, our understanding of the comparative benefits and the endogenous emergence of institutions remains limited. Here, we study the emergence and performance of sanctioning institutions in a public goods context when individuals are free to migrate between different institutions. We show experimentally that efficient peer and centralized sanctioning emerge as dominant institutions that immediately generate and maintain high levels of cooperation without much need for costly punishment. The quick establishment of high cooperation is due to both the self-selection of prosocial individuals into these institutions and the institutions' intrinsically beneficial properties. In addition, voluntary migration into the centralized sanctioning institution leads to the selection of stable prosocial leaders who refrain from antisocial punishment, while remnants of antisocial punishment still exist under peer punishment.

A.2 Introduction

Institutions are pervasive in social and economic life. They "are the humanly devised constraints that shape human interaction" which include both formal institutions such as laws and constitutions as well as informal institutions such as social norms, conventions, and taboos (North, 1990). Institutions shape economic and social incentives and are therefore of paramount importance for the economic performance of individuals, groups, companies and, perhaps, even countries. In the long run, however, institutions are themselves subject to individual and political choices, and may thus be viewed as an equilibrium outcome in a broader "game" in which different institutions compete for the support of the population.

In this paper we study the endogenous emergence of a particularly important "humanly devised constraint" - sanctioning institutions - in the context of public goods

provision. Throughout human evolution, social groups faced important public goods problems that ranged from the provision of social insurance through food sharing among hunter-gatherers and cooperation during warfare between neighboring groups to the provision of effort among coworkers who receive a group bonus in case of high profits. The experimental literature on public goods provision (e.g. Fischbacher et al., 2001) has shown that many people are willing to contribute voluntarily to public goods if others do so as well, but it is generally not possible to sustain a high level of cooperation if free-riders face no sanctions (Ostrom et al., 1992; Fehr and Gächter, 2000). Historically, peer sanctions are probably the oldest form of sanctioning that emerged among hunter-gatherers long before humans developed more centralized sanctioning institutions that involved judges and central enforcement of punishments. However, peer punishment has been shown to generate high initial costs because of coordination failure among peers and because a considerable amount of initial sanctioning is necessary to establish the credibility of the punishment threat and, in small groups, punished individuals may not necessarily respond to sanctioning in a prosocial manner (Gächter et al., 2008; Dreber et al., 2008). Peer punishment is, in particular, often associated with "antisocial punishment," i.e. when low contributors punish individuals who make above average contributions to the public good (Gächter et al., 2008).

The short and medium run inefficiency of uncoordinated peer punishment raises the question whether and how human groups are capable of avoiding the high initial costs of peer punishment. We study this question in an experimental environment in which individuals are free to sort themselves into different sanctioning institutions. Ethnographic evidence indicates that early human groups were characterized by high mobility and frequent migration in and out of existing groups (Boehm et al., 1996; Kaplan et al., 2005; Wiessner, 2005; Mathew and Boyd, 2011). Thus, allowing individuals to leave and join groups freely seems to capture an important component of social life in the early evolution of humans.[1] In our experiment, individuals can sort into four different institutions;

[1]Our set up is also related to Tiebout (1956) who argues that public goods are largely provided at the level of the local community and that consumer-voters will "vote with the feet" by moving to

within an institution, they can contribute to, and benefit from, a public good that only benefits members of the institution. For simplicity, and to have a stark contrast between individual and collective interest, it is in an individual's rational self-interest to contribute nothing to the public good when he or she faces no sanctions for free-riding, but group welfare is maximized if everybody contributes the whole endowment to the public good.

One of the available institutions is characterized by the absence of any explicit opportunity for the sanctioning of individual free-riders ("no punishment"). The second institution provides an opportunity for each group member to sanction any other group member after they have observed each of their contributions to the public good. We denote this institution as "uncoordinated peer punishment" because it does not offer any explicit possibility to coordinate the group members' contribution or punishment activities. This institution has dominated the experimental economics literature in recent years starting with Ostrom et al. (1992) and Fehr and Gächter (2000). We add a cheap-talk normative request to peer punishment in a third institution, denoted by "coordinated peer punishment," as minor communication could act as a coordination device for contributions, determine when and who should be punished, and potentially affect beliefs about others' preferences. Here, each institution member can state how much he or she thinks everyone in the institution should contribute. The average of these statements is then communicated to every institution member before the contribution decision. The rationale for the coordinated peer punishment institution is that sanctioning typically does not take place in a normative vacuum. Rather, people often sanction for a reason, i.e., they punish what they consider as normatively inappropriate behavior. It thus makes sense to allow them to express their normative views and provide them with feedback about the average view in the group. The fourth and final institution ("coordinated central punishment") maintains the cheap-talk normative request but allows for the delegation of punishment to a single (central) authority elected by the group while also socializing the cost of punishment. This type of institution is prevalent in both small-scale societies (e.g.

communities that best satisfy their preferences.

village elders and tribal chiefs, with collective punishment by the group) and large-scale societies (e.g. police, courts, and prisons funded by taxes).

Our results show that both coordinated peer punishment and centralized punishment function very well and establish extremely high cooperation levels *from the beginning with little need for sanctions*. After an initial adjustment phase, subjects thus predominantly choose these two institutions, while the other two institutions - no punishment and uncoordinated peer punishment - become depopulated. In fact, the uncoordinated peer punishment institution is almost never chosen, even at the very beginning.

The centralized punishment institution completely removes the inefficiencies of uncoordinated peer punishment and already leads to payoff levels that are significantly greater than in "no punishment" in the first period. In addition, centralized punishment removes antisocial punishment. The high efficiency of this institution is based on the two key facts. First, many prosocial individuals (i.e., those with prosocial other-regarding preferences) enter this institution at the very beginning, leading to high normative contribution requests and the selection of a prosocial central authority. Rather than being merely cheap talk, the high normative requests are associated with high actual contributions - individuals seem to use the average contribution request as a coordination device. Subjects thus quickly establish a strong cooperative culture in the centralized punishment institution. The second reason for the superiority of centralized punishment is due to its intrinsically beneficial properties - even in the absence of endogenous sorting of subjects, this institution is capable of producing high cooperation with comparably little punishment costs.

Coordinated peer punishment shares many of the good properties of centralized punishment. Many prosocial individuals immediately enter this institution; they establish very high normative requests followed by equally high contributions. However, this institution requires more actual sanctions during the first few periods, and some antisocial punishment still persists. Therefore, initial payoffs are not larger than in "no punishment" but - in contrast to the uncoordinated punishment institution (Gächter et al., 2008) - pay-

offs are never smaller than in "no punishment." In fact, coordinated peer punishment outperforms "no punishment" in terms of overall payoff after only three or four periods. Taken together, our results show that efficient punishment institutions emerge endogenously through a competitive process in an environment in which people can "vote with their feet." Prosocial individuals play a key role in this process because they quickly establish a cooperative culture that considerably shortens the length of time that it takes to render an institution efficient. While uncoordinated peer punishment incurs large initial costs, the combination of endogenous sorting of prosocial individuals with the possibility of coordinating group behavior through normative requests very quickly makes both peer punishment and centralized punishment the superior institutions.

Our results speak to a growing body of research on endogenous choice and cooperation. Broadly speaking, these papers fall into three categories: endogenous groups with fixed institutions (e.g. Ahn et al., 2008, 2009), fixed groups with endogenous institutions (e.g. Kosfeld et al., 2009; Sutter et al., 2010), and endogenous groups with endogenous institutions (e.g. Gürerk et al., 2006). The last category is the smallest but also best captures the idea of "voting with feet" to select communities that satisfy the individual's preferences due to Tiebout (1956). Our paper falls into this category.

Gürerk et al. (2006, 2011, 2013) precede our work in allowing for both endogenous groups and endogenous institutions. Their treatments restrict subjects to only two institutions, (i) a non-sanctioning institution and (ii) an uncoordinated peer-sanctioning institution. Their treatments allow for punishment, reward, or both in the peer-sanctioning institution. However, only one sanctioning institution is available in any treatment and subjects cannot express their normative requests in any of their treatments. Our study shows that the existence of these requests is not innocuous because - if available - subjects immediately leave the uncoordinated punishment institution in favor of coordinated peer punishment or centralized punishment. In addition, we provide insights into the key role of prosocial individuals for the quick and smooth functioning of punishment institutions because we also elicit an independent measure of subjects' prosociality.

Like us, Grechenig et al. (2013) extend the punishment institutions to also allow for centralized punishment. In contrast, however, our central authority is elected by institution members each period and also bears part of the cost of punishment, while they exogenously and permanently assign subjects to be the central authority. The election allows us to examine to whom subjects delegate authority, while the exogenous and permanent assignment in Grechenig et al. (2013) rules out this possibility. We also allow for a cheap-talk normative request that serves as a coordination device for contributions, which helps to quickly establish a cooperative culture. Finally, our work differs from Grechenig et al. (2013) because they neither provide an independent measure of subjects' prosociality nor do they compare the functioning of institutions under endogenous sorting and under exogenous assignment of individuals to institutions.

Our approach may also be a useful complement to research on the persistence of macro-institutions and historical development. Lab experiments are not a substitute for empirical data and identification of effects due to natural experiments and the use of instruments. However, laboratory experiments can provide a controlled setting to test theories that emerge from naturally-occurring data without the need for (non-experimenter) exogenous variation. Nunn (forthcoming) includes a discussion of mechanisms underlying the persistent effects of institutions in historical development and focuses on culture, norms, genetics, and coevolutionary processes. Much of the current knowledge of these various factors has been shaped by work involving cross-cultural lab-in-field experiments (Henrich et al., 2006, 2010; Marlowe et al., 2008).[2] Kimbrough et al. (2008) is a noteworthy example of how controlled lab experiments can help inform our understanding of historical development, in which they focus on the emergence of long-distance trade.

[2]For example, an unresolved question in the macro-institutions literature that is also relevant for development policy is whether institutions will continue to be successful when exogenously imposed in new environments. Acemoglu et al. (2011) find evidence supportive of exogenous imposition in the French Revolution, while Berkowitz et al. (2003a,b) argue that the evidence for exogenously imposed institutions following World War II and the fall of the Soviet Union has been much more mixed. Lab and field experiments can potentially identify the important common features of institutions and which specific features may be effective in similar contexts (eg. common culture) but likely to fail in alternative environments. They may also be able to do so at lower cost while improving the effectiveness of costly large-scale interventions by identifying critical aspects in advance.

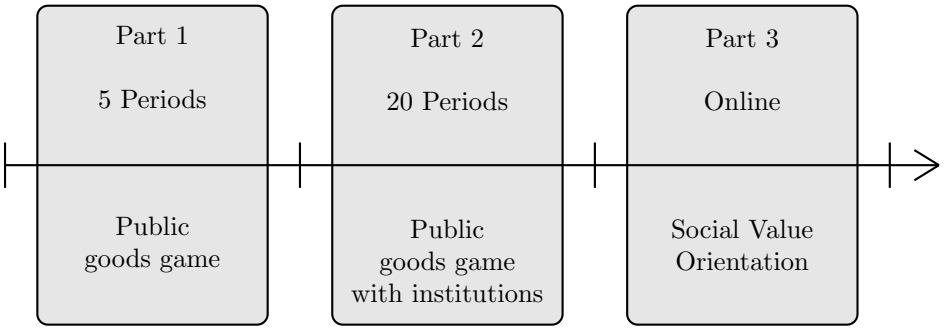


Figure A.1: Experiment timeline.

The remainder of the paper is organized as follows. Section A.3 presents our experimental design in detail. Section A.4 presents our results. Section B.8 concludes the paper and discusses open questions and possible fertile future studies.

A.3 Experimental design and procedures

The experiment consists of three parts. Parts 1 and 2 are conducted in the lab during the same session. Part 3 is conducted online after subjects leave the lab and provides an independent measure of subjects' social preferences.

Subjects are initially assigned to a large group of size $N \in \{9, 11, 12\}$ that stays fixed for Parts 1 and 2 of the experiment. We attempted to have 12 members in all groups but occasionally used smaller groups due to subjects registering for the study but failing to show up to the lab. Subjects are randomly assigned a unique identification number from the set {1,2,. . . ,N} which also stays fixed for Parts 1 and 2 of the experiment.

Part 1 consists of five periods of a typical public goods game without punishment. Part 2 consists of 20 periods of a public goods game in which subjects can endogenously form subgroups by adopting different punishment institutions; we also include a control treatment to disentangle institutional effects from selection effects (see Section A.3.2). Part 3 is conducted online and measures social preferences using the Social Value Orientation scale of Murphy et al. (2011). The experiment timeline is summarized in Figure A.1.

A.3.1 Part 1: Public goods game

Subjects begin by playing five rounds of a typical public goods game without punishment. At the beginning of each period, subjects receive an endowment of points, e, and can contribute any amount $g_i \in \{0, 1, \ldots, e\}$ to a group project. Each point contributed to the group project is multiplied by m and shared equally among all N group members. Each point not contributed to the project goes into a private account. Thus, per-period earnings are given by

$$\pi_i = e - g_i + \left(\frac{m}{N}\right) \sum_{j=1}^{N} g_j. \tag{A.1}$$

A social dilemma exists whenever (i) $\partial \pi_i / \partial g_i = (m/N) - 1 < 0$ and (ii) $\partial (\sum_{j=1}^{N} \pi_j) / \partial g_i = m - 1 > 0$ for all $g_i > 0$. Condition (i) means that own payoff is decreasing in own contribution to the project, so that free-riding is individually optimal for payoff-maximizers. Condition (ii) means that aggregate payoffs are increasing in own contribution to the project, so that full contributions by all group members are socially optimal.

The marginal per-capita return (MPCR) is given by (m/N) and is decreasing in group size N. We set $m = 1.5$ in the experiment, so that MPCR ranged from 0.125 when $N = 12$ to 0.167 when $N = 9$. Previous experiments suggest that these values for the MPCR and group size should lead to the breakdown of cooperation within five periods.[3] Subjects should therefore directly experience the public goods problem during Part 1 of the experiment and have a potential motivation to form institutions to establish and maintain cooperation in order to improve their own payoffs. In each period and after making private contribution decisions, group members are informed of every group member's contribution. In addition to $m = 1.5$, we set $e = 20$ and $m = 1.5$. The parameter values are summarized in Table A.1.

[3]Hamman et al. (2011) used an MPCR of 0.15 with fixed group size of 9. Average contributions began around 45% of the endowment in the first period and declined to about 15% in period 5.

Parameter	Value	Meaning
e	20	Endowment
m	1.5	Multiplier for contribution to public good
N	9, 11, or 12	Group size

Table A.1: Parameter values used in Part 1.

A.3.2 Part 2: Public goods game with endogenous punishment institutions

Subjects remain in the same group and retain the same identification number in Part 2 of the experiment. Part 2 lasts for 20 periods and provides subjects with the opportunity to form institutions with other group members after experiencing the public goods problem during Part 1. At the beginning of each period, subjects individually select into one of four institutions and interact only with other group members who adopt the same institution in that period. Migration between institutions is costless, and subjects can adopt any of the institutions at the start of each period. The institutions are (i) No Punishment, (ii) Uncoordinated Peer Punishment, (iii) Coordinated Peer Punishment, and (iv) Coordinated Central Punishment. By "Coordinated," we mean the presence of a normative request that can be used as a coordination device for contributions; it does not refer to the coordination of punishment. These institutions are described in more detail in Section A.3.2.

Each period contains a contribution stage which is identical to the decision in Part 1, except that contributions to the group project only affect members of the group who adopt the same institution. In addition, a punishment stage is added which provides an additional endowment in each period. In the event that only one subject adopts a particular institution in the period, both the contribution stage and punishment stage endowments go directly to the private account, and the subject is not able to contribute to a group account. This design feature was included because the idea of a public good necessarily involves more than one person benefitting from contributions.

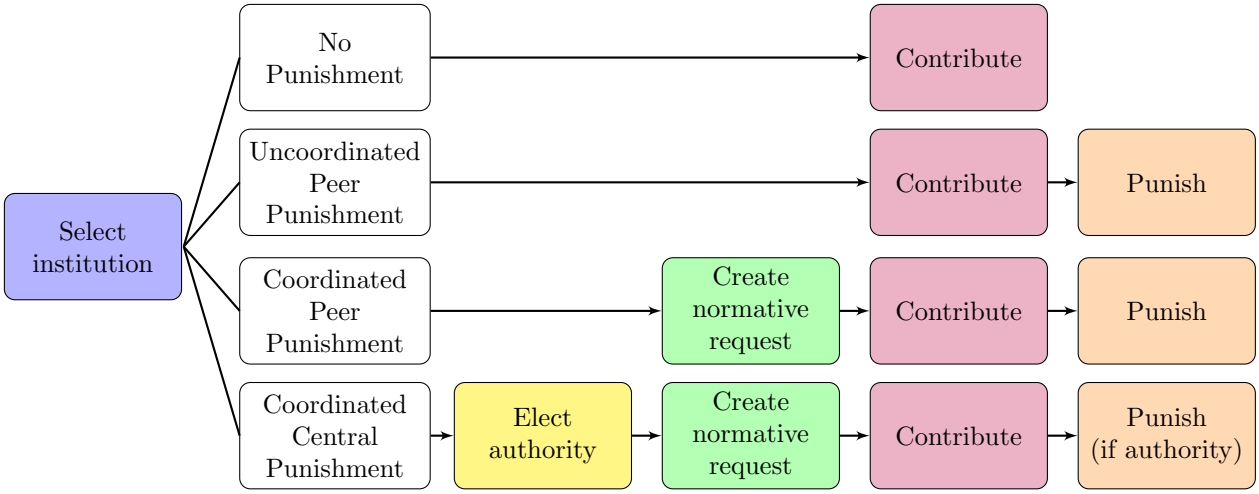


Figure A.2: Sequence of decisions in Part 2.

We next describe the four institutions in greater detail. During each period, subjects make several decisions, and their choice opportunities depend partly on the institution adopted. We then describe the information provided to subjects at each decision point. Finally, we describe the material payoffs resulting from subjects' choices.

Institutions

Subjects begin each period by selecting which institution they want to adopt in the current period. For the remainder of the period, subjects only interact with other group members who have also adopted the same institution. The sequence of decisions made during each period is summarized in Figure A.2.

No Punishment. No Punishment is identical to Part 1, except that (i) contributions to the group project only affect members of the group who adopt the No Punishment institution and (ii) subjects receive a second endowment in the punishment stage which is added directly to their earnings.

Uncoordinated Peer Punishment. In Uncoordinated Peer Punishment, subjects make the same contribution decision as in No Punishment. However, during the punishment stage, subjects can assign deduction points to other institution members which

reduce the recipient's earnings at a cost to the person assigning punishment. The cost structure for punishment is described in Section A.3.2 and defined in equation (A.4).

Coordinated Peer Punishment. The contribution and punishment stages in Coordinated Peer Punishment are identical to Uncoordinated Peer Punishment. However, prior to the contribution stage, each institution member makes a normative request by privately answering the question, *"How many points do you think each participant should contribute to the project?"* The average of these private responses is reported on each institution member's computer screen during the contribution stage.[4] The normative request is non-binding and public knowledge within the institution. The cost structure for punishment is described in Section A.3.2 and defined in equation (A.4).

Coordinated Central Punishment In Coordinated Central Punishment, one member of the institution is elected to assign all of the punishment for the group, and the total cost of punishment is shared equally by all institution members. At the start of the period, subjects in Coordinated Central Punishment vote for a single institution member to assign the punishment; the central authority is the person who receives the most votes, and ties are broken randomly. After casting their votes, subjects then create a normative request in the same manner as in Coordinated Peer Punishment. Then, subjects enter the contribution stage, where they are informed of the normative request and make the same contribution decision as in all other institutions. Finally, during the punishment stage, only the central authority can assign deduction points to institution members, and these deduction points reduce the recipient's earnings at a cost which is shared equally by all members of the institution.[5] The cost structure for punishment is described in

[4] While the normative request is cheap talk, it may act as a coordination device for equilibrium selection when multiple cooperative equilibria exist; social preference models, eg. Fehr and Schmidt (1999), often suffer from the problem of multiple equilibria. In such cases, the normative request is both self-signalling and self-committing in the sense of Farrell and Rabin (1996).

[5] To our knowledge, this particular institution is novel. Therefore, in Appendix A.7.1, we characterize a class of cooperative equilibria in which some individuals have inequity averse social preferences in a one-shot interaction; we do so for comparison with the other institutions based on previous research, eg. Fehr and Schmidt (1999). The existence of only one group member who is strongly averse to inequality is both necessary and sufficient for the existence of cooperative equilibria. Under peer punishment with

Section A.3.2 and defined in equation (A.5).

Other possible institutions. Our design superficially looks to be a 3×2 factorial design with two omitted institutions, (i) Coordinated No Punishment and (ii) Uncoordinated Central Punishment. We first want to stress that we do not actually have a factorial design, as we do not run separate sessions or treatments for each institution. Instead, we allow subjects to endogenously determine – based on their decisions – whether all institutions, some institutions, or only one institution will be adopted in each period. Our motivation is to understand what institutions people will adopt in natural environments and to see how successful these institutions become in establishing and maintaining cooperation.

In our view, these two potential institutions are not relevant to understanding natural environments. We start from the perspective that peer punishment is always available in natural environments, negating the need to include Coordinated No Punishment; in addition, numerical cheap-talk, such as the normative request we use in this paper, typically is not effective in establishing cooperation.[6] The No Punishment institution is a convenient benchmark and has been the traditional way of examining social dilemmas, hence its inclusion. We also do not find it plausible that a group would somehow lose the ability to establish a normative request when moving from peer punishment to centralized punishment, which negates the need to include Central Punishment without a normative request.

We also face a practical concern regarding the number of institutions because we do not allow a subject to contribute to a public good if she is the only one to adopt

social preferences, a single group member who is inequity averse is necessary but not always sufficient for the existence of cooperative equilibria.

[6] Verbal and written communication often enhances cooperation (Isaac and Walker, 1988a; Sally, 1995; Ostrom, 1998); however, numerical communication in which written messages cannot be sent is generally unable to establish cooperation and occasionally performs worse than an environment without communication (Wilson and Sell, 1997; Bochet et al., 2006; Bochet and Putterman, 2009). In light of these previous findings, one can reasonably assume that numerical communication itself is not the driving force behind cooperation in our institutions with a cheap-talk normative request, though it may interact with and enhance institution performance.

the institution in the period. If we increase the number of institutions, subjects face a primary concern of adopting an institution based on their beliefs that at least one other person will adopt the same institution; otherwise, they will not be able to even potentially benefit from a public good. Concerns about beliefs in such cases casts doubt on any inferences that can made regarding preferences over institutions, as beliefs about others' choices can override one's own preference. These concerns also led us to omit the two implausible institutions.

Information

The information provided during each period is summarized in Figure A.3. All information described in Figure A.3 is provided in each institution even if no decision is made at that point. We intentionally chose this feature to rule out desire for increased information as a confounding explanation for institution selection. Subjects receive more information about their own institution than they do about the other institutions, which captures the notion that we a great deal about the people we interact with but have only limited information about those with whom we do not interact.

At the beginning of each period and before subjects join an institution, all subjects are informed of the number of group members who joined each institution in the previous period and the average earnings for each institution in the previous period. After subjects join an institution, they learn the contribution of each current institution member in the previous period; the Coordinated Central Punishment institution elects the central authority at this point.[7] During the contribution stage, the Coordinated Peer Punishment and Coordinated Central Punishment institutions are informed of the institution's normative request. Finally, at the punishment stage, members of all institutions observe the contribution of each institution member in the current period.

[7]In the first period of Part 2 (period 6 overall), subjects are informed of the average contribution of each current institution member during Part 1.

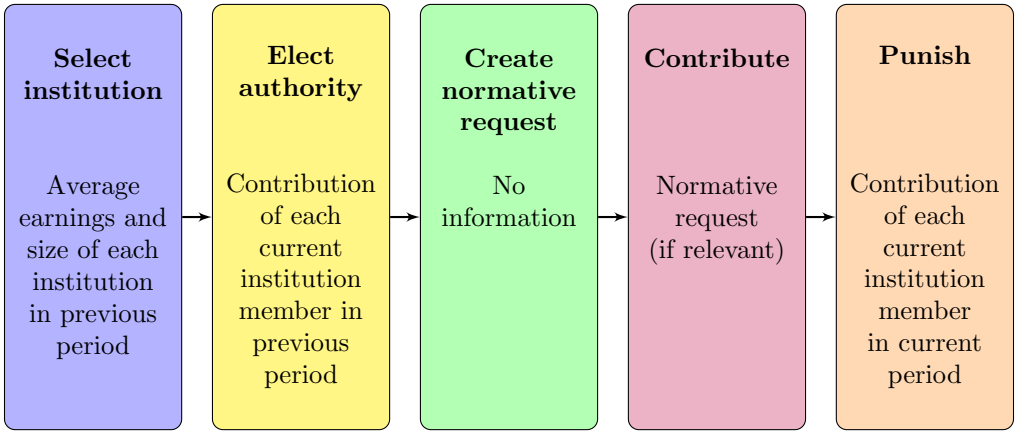


Figure A.3: Sequence of information in Part 2. Information is known by all institution members regardless of whether a decision is made at that stage.

Material payoffs

Earnings in each period are the sum of the contribution stage earnings and the punishment stage earnings. Contribution stage earnings for individual i in institution $inst$, $\pi^1_{i,inst}$, are given by

$$\pi^1_{i,inst} = e^1 - g_{i,inst} + \left(\frac{m}{s_{inst}}\right) \sum_{j=1}^{s_{inst}} g_{j,inst}. \tag{A.2}$$

where e^1 is the contribution stage endowment, $g_{i,inst}$ is the individual's contribution to the institution project, m is the multiplier for contributions to the institution project, and s_{inst} is the endogenously determined institution size (number of institution members).

Punishment stage earnings for individual i in institution $inst$, $\pi^2_{i,inst}$, are given by

$$\pi^2_{i,inst} = e^2 - c_{i,inst}(d) - r \sum_{j=1}^{s_{inst}} d_{ji,inst} \tag{A.3}$$

where e^2 is the punishment stage endowment, $c_{i,inst}(d)$ is the cost of assigned punishment, r is the reduction in earnings for each deduction point received, and $d_{ji,inst}$ is the number of deduction points assigned to subject i by subject j in institution $inst$. The cost of assigned punishment, $c_{i,inst}(d)$, differs by institution and is given by

$$c_{i,inst}(d) = \sum_{j=1}^{s_{inst}} d_{ij,inst} \tag{A.4}$$

Parameter	Value	Meaning
e^1	20	Endowment for contribution stage
m	1.5	Multiplier for contribution to public good
s_{inst}	Endogenous	Institution size
e^2	20	Endowment for punishment stage
r	3	Reduction in earnings from unit of punishment
$c_{i,inst}(d)$	See equations (A.4) and (A.5)	Cost of assigned punishment

Table A.2: Parameter values used in Part 2.

if $inst \in$ {Uncoordinated Peer Punishment, Coordinated Peer Punishment} and

$$c_{i,inst}(d) = \frac{\sum_{j=1}^{s_{inst}} d_{Aj,inst}}{s_{inst}} \tag{A.5}$$

if $inst =$ Coordinated Central Punishment, where $d_{Aj,inst}$ is the number of deduction points assigned to subject j by the central authority A in the Coordinated Central Punishment institution. In the No Punishment institution, $c_{i,inst}$ and $d_{ji,inst}$ are always equal to zero for all i and j.

Finally, we imposed a bankruptcy condition so that earnings in a single period could not be negative. Subjects would still have to pay for assigned deduction tokens even if the tokens could not reduce the recipient's earnings any further. Therefore, per-period earnings are given by

$$\pi_{i,inst} = \max\{\pi^1_{i,inst} + \pi^2_{i,inst}, 0\}. \tag{A.6}$$

Table A.2 summarizes the parameter values used in the experiment. We set $e^1 = e^2 = 20$, $m = 1.5$, and $r = 3$. The total cost of each deduction point is 1, but the cost for each institution member is determined according to equation (A.4) in institutions utilizing peer punishment and equation (A.5) in the central punishment institution.

Control treatment – Exogenously assigned perfect matching groups

Under endogenous institution selection, institutions may be successful in establishing and maintaining cooperation for two primary reasons. First, cooperative individuals may select into the same institutions, and cooperation is likely to follow regardless of the institution itself. Second, the institution may create incentives that induce cooperative behavior, regardless of whether the individuals joining the institution are generally cooperative. In our view, the most likely explanation is an interaction of these aspects. Our design with endogenous selection does not allow us to disentangle these explanations. Therefore, we conducted a control treatment with exogenous assignment. For each group in our endogenous selection sessions, we create a matching group of the same size and with identical migration patterns. This exogenous assignment allows us to examine the effects of institutions independently of self-selection.

A.3.3 Part 3: Social Value Orientation

Part 3 of the experiment consists of the Social Value Orientation (SVO) measure of Murphy et al. (2011) which consists of six allocation decisions between oneself and one other anonymous individual. The allocation decisions constitute modified versions of the dictator game (Forsythe et al., 1994) in which the relative price of giving varies, similar to the approach in Andreoni and Miller (2002); unlike Andreoni and Miller (2002), the SVO measure provides a numeric score which can be used to make comparisons across individuals. Higher SVO scores indicate greater prosociality. Full details of the SVO measure are provided in Appendix A.7.2. This part of the experiment was conducted online after subjects completed Parts 1 and 2 in the lab. Subjects knew that one of their allocation decisions would be selected and implemented and that they would be the recipient of a different person's allocation decision; payments were mailed to subjects.

A.3.4 Experimental procedures

Sessions for Part 1 and Part 2 were conducted in a computer lab at the University of Zurich in December 2012 and April 2013. Part 3 was conducted online using Qualtrics (www.qualtrics.com). Experimental instructions are provided in the Appendix. Subjects were mostly students from the University of Zurich and Swiss Federal Institute of Technology (ETH-Zurich). Recruitment was conducted using ORSEE (Greiner, 2004), and we excluded students who listed economics or psychology as their major in ORSEE from receiving invitations. Experiments were programmed in z-Tree (Fischbacher, 2007).

Points were used as the experimental currency and converted to Swiss Francs (CHF) at the end of the study; subjects were informed of the exchange rate in the instructions. In Parts 1 and 2, conducted during the same lab session, the exchange rate is 1 point = CHF 0.05, and final earnings were the sum of all per-period earnings. In Part 3, conducted online, the exchange rate is 1 point = CHF 0.10. Average earnings were CHF 56.41 for the lab session (consisting of both Parts 1 and 2), including a show-up fee of CHF 10. Earning from Part 3 were CHF 15-20. Lab sessions lasted 2.5-3 hours on average, and the online portion of the experiment took 10-20 minutes.

Overall, 256 subjects participated in the lab sessions; 128 subjects participated in the endogenous selection treatment, and another 128 subjects participated in the exogenous assignment treatment. Each treatment consisted of eleven groups in total. Nine of these groups had twelve members. One group of nine members and one group of eleven members were used in each treatment due to subjects not coming to the lab. Partner matching was used in Parts 1 and 2, so each group remained fixed for the entire lab session.

For subjects in the endogenous treatments (where we predict assortment effects), 84 of 128 subjects (66%) completed Part 3 of the experiment online. We do not find obvious evidence for selection effects. Subjects who completed Part 3 (average age = 22.0, percent female = 0.37, average earnings in Parts 1 and 2 = CHF 57.40) are similar in observable characteristics to those who did not complete Part 3 (average age = 22.3, percent female = 0.34, average earnings in Parts 1 and 2 = CHF 56.66).

A.4 Results

We begin by taking a revealed preference approach to institutions. When subjects are given a choice, which institutions are adopted? The answer provides our first major result and demonstrates that Uncoordinated Peer Punishment is almost never adopted. We include the institution in Result A.1 but drop it from subsequent analysis due to lack of observations.

Result A.1. *Subjects are initially evenly distributed between the No Punishment (NP), Coordinated Peer Punishment (PP), and Centralized Punishment (CP) institutions while Uncoordinated Peer Punishment is almost never chosen. Over time, only Coordinated Peer Punishment and Centralized Punishment survive.*

Evidence for Result 1. The result is documented in Figure A.4, which displays the aggregate distribution of subjects (i.e. data from all groups is pooled to calculate the distribution). The figure shows that initially a substantial number of subjects are in NP, but after period 5, the percentage of subjects entering NP becomes negligible. In addition, the figure shows that, from the very beginning, the share of subjects selecting Uncoordinated Peer Punishment is negligible.

The strong dominance of coordinated peer sanctioning and centralized sanctioning raises the question of why these two institutions prevail. With a revealed preference approach, one would suspect these institutions to prevail because they are successful either in maintaining cooperation or increasing earnings. We find evidence for both conjectures.

Result A.2. *Under both PP and CP, very high cooperation levels are quickly obtained. Initially, only the CP institution outperforms the NP institution in terms of aggregate payoff. However, within a few periods, PP also outperforms the NP institution.*

Evidence for Result 2. The trend for contributions can be seen in Figure A.5a. The unit of observation is the matching group, so that the average contribution for an institution is first taken at the group level; the result is then averaged across groups, which

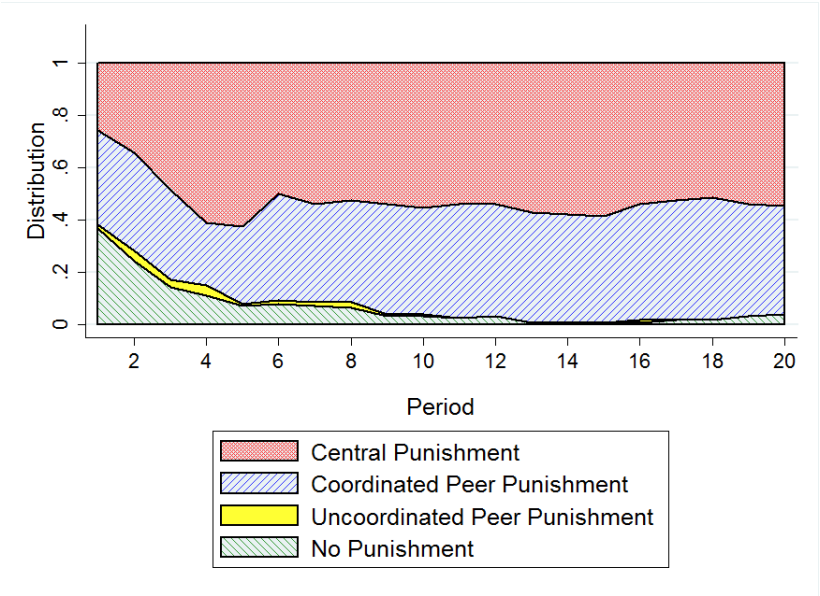
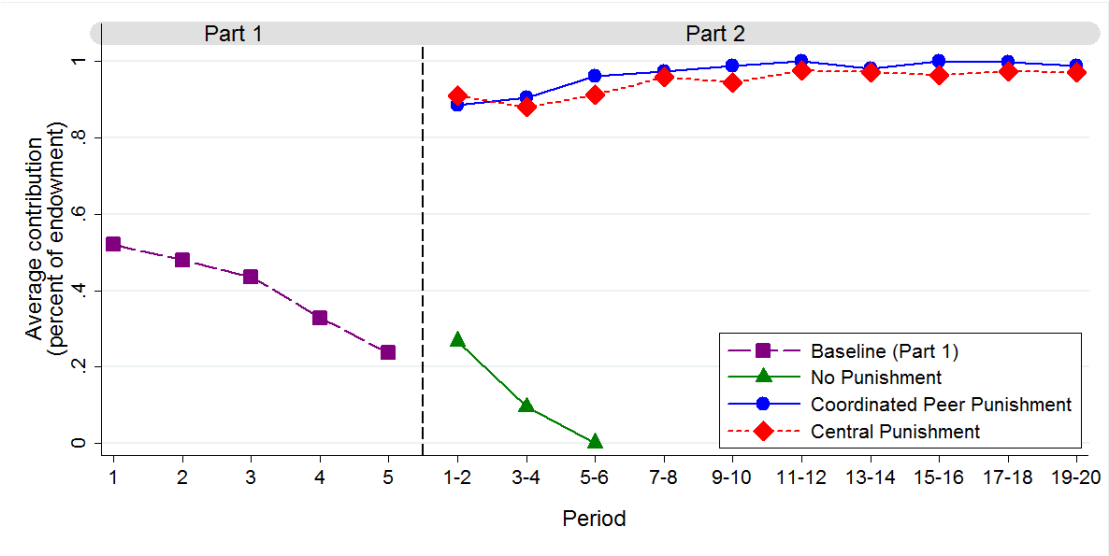


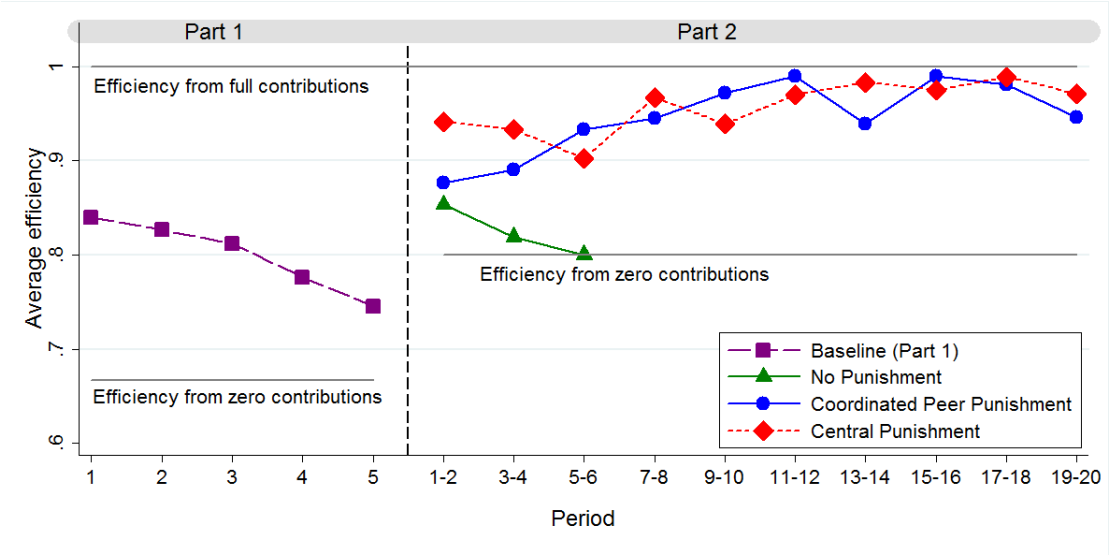
Figure A.4: Distribution of subjects across institutions in Part 2.

is displayed in the figure. Figure A.5a shows that already in the first period of Part 2 the average contribution (as a percentage of the endowment) in PP and CP is about 90% and soon reaches close to 100%. In contrast, average contributions in NP quickly decline to low levels. Evidence for earnings is provided in terms of efficiency, which we define as the percentage of the social optimum (full contributions by all subjects) earned by subjects in the institution. Average efficiency can be seen in Figure A.5b. In the first period of Part 2, efficiency is significantly higher in CP relative to NP (Mann-Whitney U, p=0.001). However, efficiency in PP is only marginally significantly different than in NP during the initial period (Mann-Whitney U, p=0.066). By period 4, the efficiency in PP is significantly higher than in the NP.

Why do coordinated peer punishment and centralized punishment perform so well in terms of quickly establishing and maintaining cooperation and in terms of aggregate payoffs? One reason could be that the institutions themselves have beneficial intrinsic properties that lead to high performance – at least in the long run – regardless of the migration pattern (i.e. even with random assignment to institutions). A second reason could be that prosocial individuals are the first ones to migrate into these institutions and quickly establish a beneficial social norm of full cooperation such that whose who



(a) Average contribution.



(b) Average efficiency.

Figure A.5: Contributions and efficiency under endogenous selection. The unit of observation is the matching group, so that the average for an institution is first taken at the group level; the result is then averaged across groups, which is displayed in the figure. Efficiency is the percentage of the socially optimal outcome (full contributions) earned by subjects. Efficiency from zero contributions differs between periods 1–5 (Part 1) and periods 6-25 (Part 2) due to the additional endowment received in periods 6–25.

join later can be smoothly integrated into a "high cooperation society." In the following, we study these two potential factors.

Result A.3. *Under exogenous assignment, institutions that allow for sanctioning and normative requests induce cooperative behavior. These institutions perform better in the*

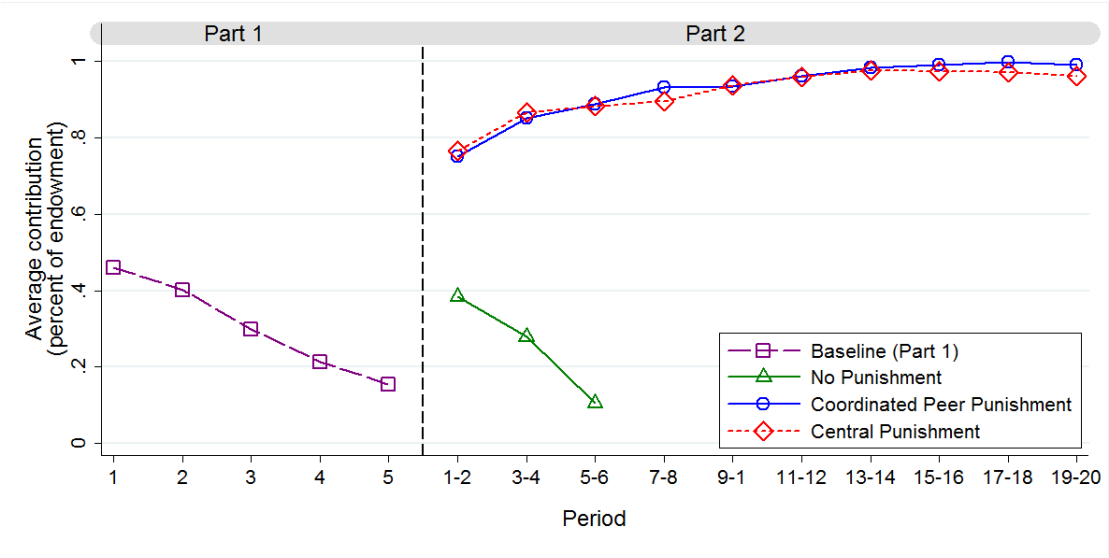
long run than NP under exogenous assignment. Under both endogenous selection and exogenous assignment, the normative request is an effective coordination device for contributions, and contributions increase immediately upon entering the sanctioning regimes from the non-sanctioning institution.

For the sake of clarity, we examine the evidence for each of these properties separately. We first examine the effectiveness of PP and CP under exogenous assignment. Selection effects cannot explain superior performance of these institutions because subjects are exogenously assigned to an institution.

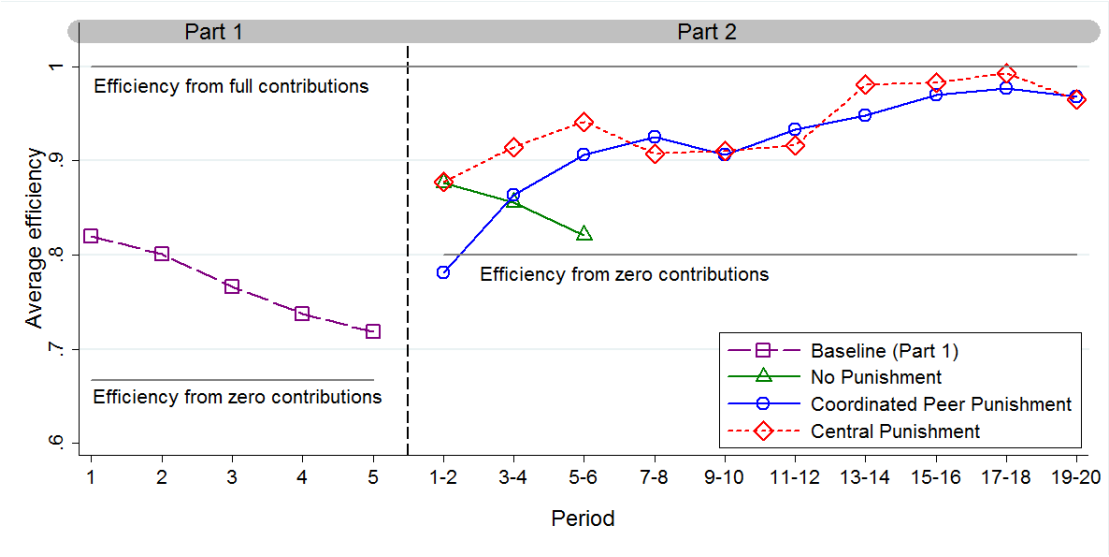
Result A.3a. *Institutions that allow for sanctioning and normative requests perform better than NP after a few periods. Initially, the efficiency in PP and CP is not higher than in NP; however, after a few periods, both PP and CP outperform NP.*

Evidence for Result 3a. The trend for contributions can be seen in Figure A.6a. Figure A.6a shows that average contributions as a percentage of the endowment are roughly 75% already in the first period of Part 2, and they reach 100% after 12 periods. Thus, contributions are significantly higher in the punishment institutions than in NP from the beginning for both PP (Mann-Whitney U, p=0.066) and CP (Mann-Whitney U, p=0.001). Average efficiency is documented in Figure A.6b, which shows that, during the first few periods of Part 2, the efficiency in the NP treatment is rather similar compared to the PP and the CP; however, from periods 5-6 onwards, efficiency in both PP and CP is significantly higher than in NP. These results indicate that under exogenous assignment there are considerable initial efficiency losses due the punishment activities of the players.

The aggregate pattern of contributions and efficiency in our coordinated peer punishment institution under exogenous assignment is remarkably similar to previous studies on uncoordinated peer punishment, in which the institution induces high contributions but initially performs worse than a sanction-free environment due to the use of punishment (Gächter et al., 2008). The existence of a strong institutional effect on contribution levels



(a) Average contribution.



(b) Average efficiency.

Figure A.6: Contributions and efficiency under exogenous assignment. The unit of observation is the matching group, so that the average for an institution is first taken at the group level; the result is then averaged across groups, which is displayed in the figure. Efficiency is the percentage of the socially optimal outcome (full contributions) earned by subjects. Efficiency from zero contributions differs between periods 1–5 (Part 1) and periods 6-25 (Part 2) due to the additional endowment received in periods 6–25.

suggests that we should also observe changes in contribution behavior when individuals migrate into PP and CP from NP. Indeed, we do observe these changes.

Result A.3b. *Individual contributions never decrease when individuals migrate from NP to either PP or CP, and most people increase their contributions upon entering a*

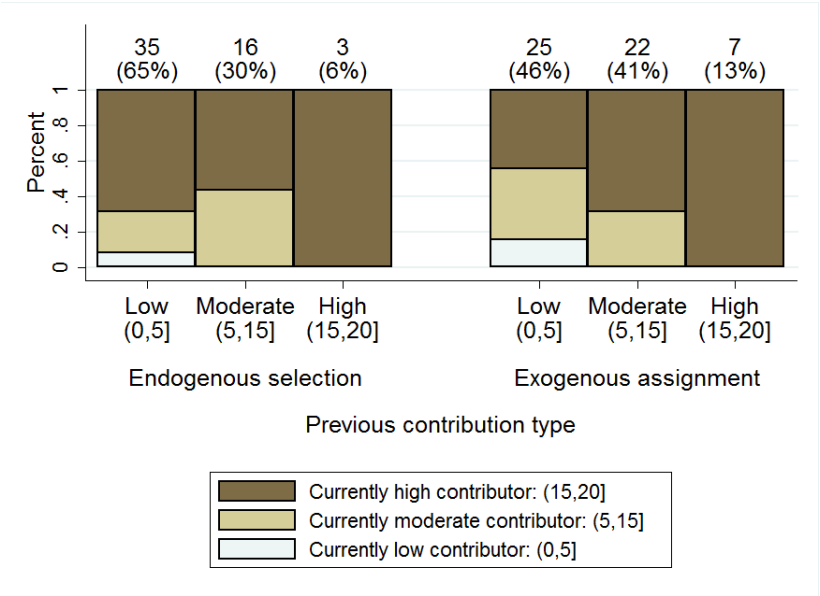


Figure A.7: Contribution behavior immediately before and after entering a punishment institution. Horizontal axis indicates the contribution type in the No Punishment institution in the last period before joining a punishment institution. Bar height indicates relative frequency of contribution type upon entering a punishment institution conditional on type before migrating. Number above bars indicates the overall number of subjects from that treatment who fall into the category.

sanctioning regime.

Evidence for Result 3b. Figure A.7 displays subjects' contribution types upon entering a punishment institution as a function of contribution type in the period immediately before entering the punishment institution. For example, approximately 70% of the subjects who were low contributors in endogenous selection (far left in Figure A.7) become high contributors upon entering a punishment institution. Taken together, Figure A.7 documents that even many of those subjects who were low contributors before they entered a punishment institution immediately turned into higher contributors upon entering a punishment institution.

We acknowledge that while this behavioral change provides evidence for institutional effects, the result itself is not terribly surprising. It is well documented in the literature that institutions allowing for punishment induce changes in behavior, even within the same subject (Fehr and Gächter, 2000; Gürerk et al., 2011). The noteworthy aspect of

the institutions adopted by subjects is the use of the normative request. While the threat of punishment may increase contributions, there is still no way of knowing the "right" contribution and what others expect you to give. The normative request, despite being cheap talk, solves this problem, even though the request itself is not generally effective in the absence of punishment.[8]

Result A.3c. *The normative request is an effective coordination device for contributions in PP and CP, and it is effective under both endogenous selection and exogenous assignment.*

Evidence for Result 3c. We restrict attention to the first five periods of Part 2 because average contributions tend towards full contributions quickly and eliminate any variation in the data afterwards. To provide support for Result A.3c, we regress subjects' contribution levels on the average normative request in a group and a constant and a restricted model in which we set the coefficient on the constant term to zero (i.e. $\beta_0 = 0$); these models are reported in Table A.3.[9] We can only reject the restricted model in favor of a model with both normative request and constant for CP under endogenous selection. However, for commonly observed values of the normative request (16-20), the predicted contributions almost perfectly match the normative request.[10] For the remaining condi-

[8]See footnote 6 for discussion and references on numerical communication in public goods experiments.

[9]An econometric issue arises when running regressions on our data, but there is a simple solution. We need to cluster standard errors at the level of the group, and we have a relatively small number of groups. It is well known that in such cases the standard errors will be inconsistent and lead to over-rejection of the null hypothesis (Bertrand et al., 2004). Bootstrapping can overcome this problem, and we use a pairs cluster bootstrap-t with cluster robust standard errors proposed by Cameron et al. (2008) which performs quite well in their simulations. In this bootstrap, resampling is at the level of clusters instead of individual observations. The Wald statistics from the bootstrap samples are then used to create the distribution against which the Wald statistic from the original data is tested. Cameron et al. (2008) suggests using a wild cluster bootstrap-t procedure in OLS regressions; however, in their simulations, the pairs cluster bootstrap-t with cluster robust standard errors performs only slightly worse than the wild cluster bootstrap-t. We opt to use the pairs cluster bootstrap-t with cluster robust standard errors for consistency with later analyses in which the wild cluster bootstrap-t is inappropriate, specifically the Tobit regressions in Table A.4. The wild bootstrap involves estimating $\hat{y}_i$ and either adding or subtracting the residual ε_i with equal probability to create new pseudo-samples, where $\hat{y}_i^*$ is the value of y_i in the wild bootstrap pseudo-sample. With corner solutions in a Tobit regression, we would need to set $\hat{y}_i^* = 0$ whenever $\hat{y}_i^* < 0$. In doing so, we would have $\hat{\beta}_i^* \neq \hat{\beta}_i$, which appears inappropriate for inference using the wild bootstrap.

[10]For example, with a normative request of 16, the predicted contribution is 16.669 ($= 5.693 + 16 \times 0.686$). With a normative request of 20, the predicted contribution is 19.413 ($= 5.693 + 20 \times 0.686$).

	Endogenous Selection		Exogenous Assignment	
	Peer Punishment	Central Punishment	Peer Punishment	Central Punishment
Panel A – Restricted model (constant constrained to $\beta_0 = 0$)				
Normative request	0.981 (0.000)***	0.997 (0.000)***	1.001 (0.000)***	0.968 (0.000)***
Panel B – Unrestricted model				
Normative request	1.078 (0.000)***	0.686 (0.000)***	1.144 (0.000)***	1.073 (0.000)***
Constant	-1.827 (0.083)*	5.693 (0.581)	-2.253 (0.001)***	-1.803 (0.000)***
Likelihood ratio, $\chi^2(1)$	0.54 (0.461)	6.69 (0.010)***	2.49 (0.115)	1.69 (0.194)
N	201	292	201	292
Number of clusters	11	11	11	11
Bootstrap samples	9,999	9,999	9,999	9,999

Table A.3: The role of normative requests as a coordination device for contributions in the first five periods (periods 6-10 overall). OLS regressions with robust standard errors clustered by matching group. Bootstrapped p-values given in parentheses based on 9,999 bootstrap samples computed using pairs cluster bootstrap-t with clustered standard errors.

tions, we cannot reject the model with only the normative request as a regressor using a likelihood ratio test. In the restricted model, coefficients on the normative request range from 0.968 to 1.001, indicating that normative request is a clear coordination device leading to contributions that are not significantly different from the normative request.

We have so far provided evidence that coordinated peer punishment and centralized punishment institutions have beneficial intrinsic properties that lead to high performance – at least in the long run – regardless of the migration pattern (i.e. even with random assignment to institutions). However, these intrinsic properties were one of the two potential reasons for why these institutions perform so well in terms of quickly establishing and maintaining cooperation and in terms of aggregate payoffs. A second reason could be that prosocial individuals are the first ones to migrate into these institutions and quickly

establish a beneficial social norm of full cooperation such that those who join later can be smoothly integrated into a "high cooperation society." In the next two results, we study the latter potential factor and find evidence in support of sorting effects.

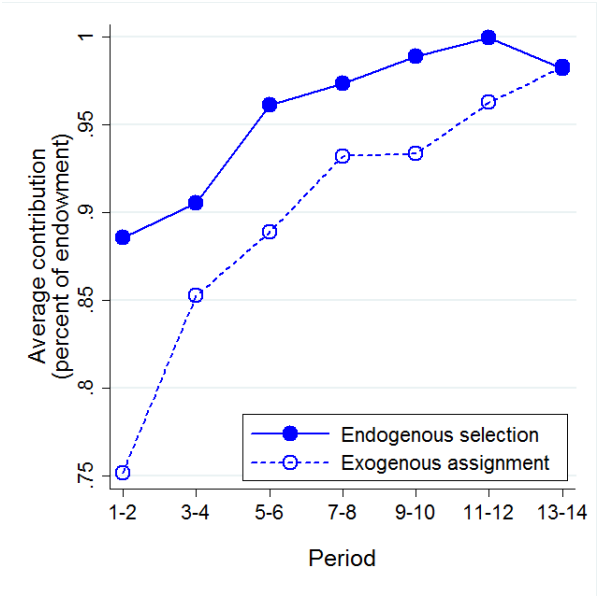
Result A.4. *Under endogenous selection, prosocial individuals are quick to migrate into the coordinated peer punishment and centralized punishment institutions and establish a culture of high cooperation. This culture of cooperation includes making higher normative requests and following through by making higher contributions than the subjects under exogenous assignment.*

Evidence for Result 4. We provide four pieces of evidence in support of the result. Statistical results presented as evidence are based on Mann-Whitney U tests.

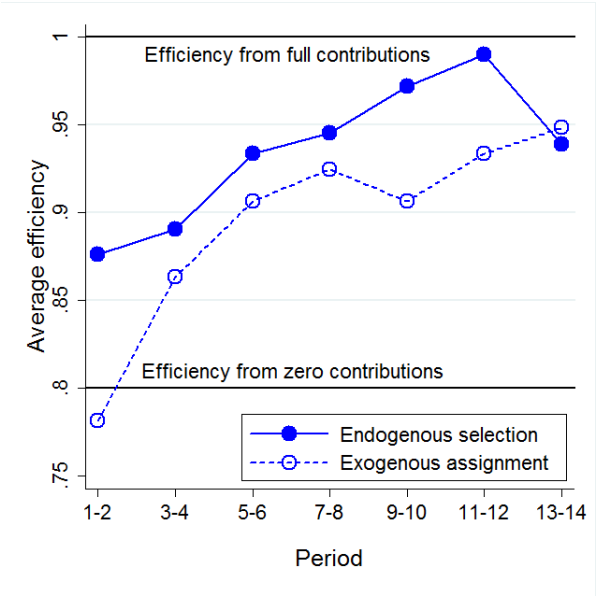
1. Contributions and efficiency are much higher in PP and CP under endogenous selection relative to exogenous assignment. While the institutions under exogenous assignment eventually converge to the performance under endogenous selection, it takes much longer. The trend across time can be seen in Figure A.8. In the first period of Part 2, the endogenously selected institutions perform significantly better with respect to both contributions (PP, p=0.021; CP, p=0.012) and efficiency (PP, p=0.016; CP, p=0.035). The unit of observation is the matching group, so that one observation is obtained for each institution in a matching group (if it is adopted).

2. Higher average contributors during the baseline public goods game in Part 1 of the experiment are more likely to adopt coordinated peer punishment and centralized punishment institutions in the first period of endogenous selection. This behavior can be seen clearly in Figure A.9a, where the average individual contribution during the baseline public goods game without punishment in Part 1 is the unit of analysis (PP > NP, p=0.012; CP > NP, p=0.015).

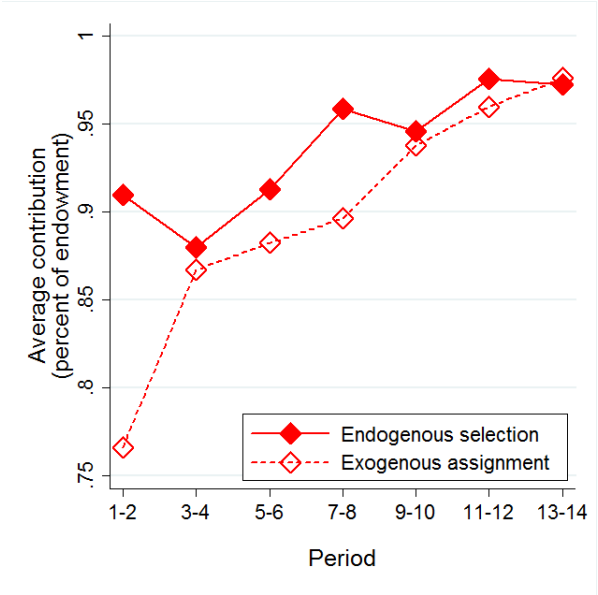
3. Prosocial individuals, as measured by the Social Value Orientation scale in Part 3 of the experiment (conducted online), are more likely to adopt coordinated peer



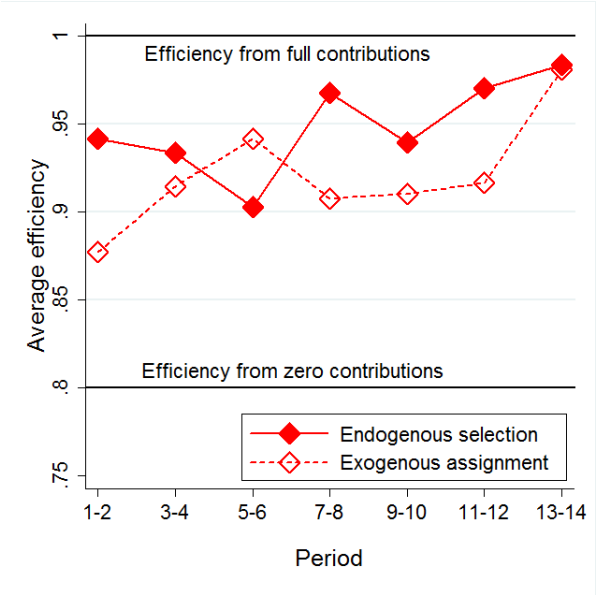
(a) Contributions in coordinated peer punishment.



(b) Efficiency in coordinated peer punishment.

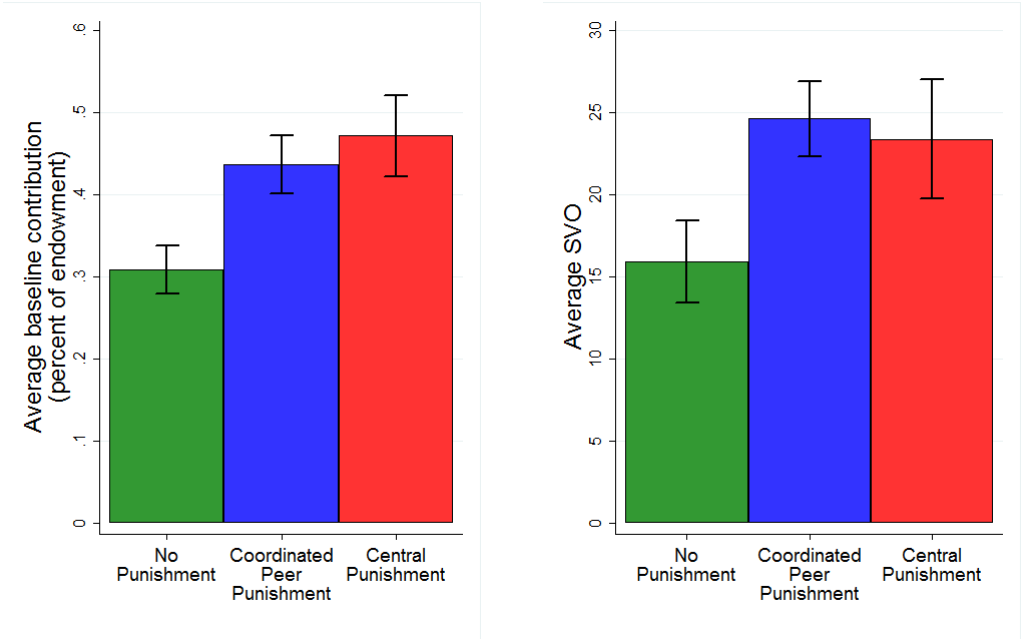


(c) Contributions in central punishment.



(d) Efficiency in central punishment.

Figure A.8: Contributions and efficiency between treatments in Part 2. The unit of observation is the matching group, so that the average for an institution is first taken at the group level; the result is then averaged across groups, which is displayed in the figure. Efficiency is the percentage of the socially optimal outcome (full contributions) earned by subjects. Periods 15-20 are omitted from the figure since there is no observable difference between treatments for either institution in these periods.



(a) Average baseline contribution

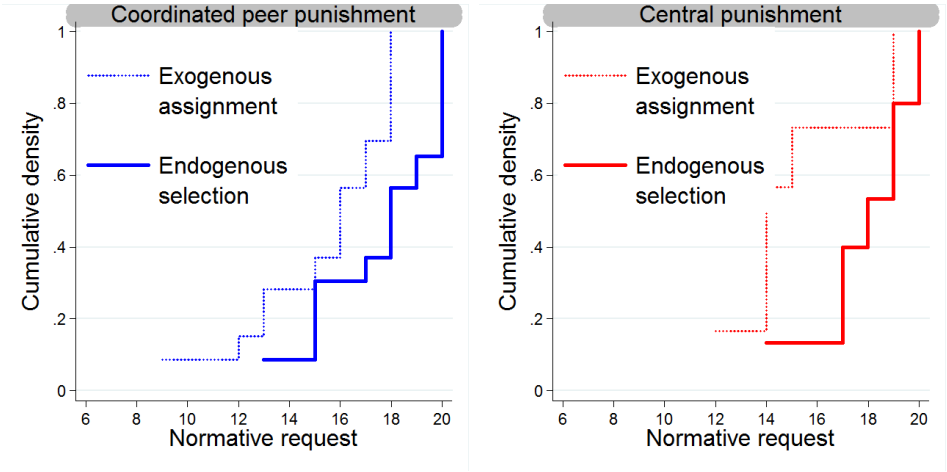
(b) Social Value Orientation

Figure A.9: Assortment in first period of endogenous selection (period 6 overall). Error bars represent standard error of the mean.

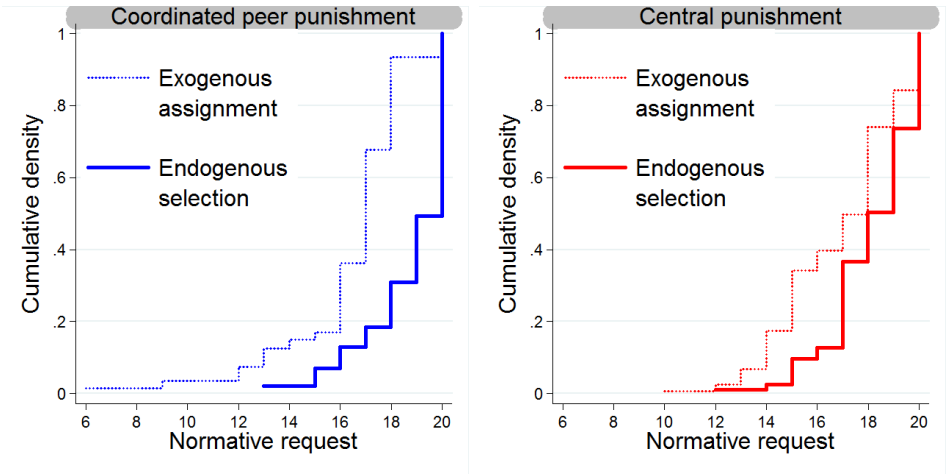
punishment and centralized punishment institutions in the first period of endogenous selection. This behavior can be seen in Figure A.9b. Individual SVO scores are the units of analysis. Since not all subjects completed Part 3 of the experiment online, we pool the PP and CP institutions for the statistical test and find that the prosociality of individuals selecting into punishment institutions is significantly higher than those who select into NP (PP and CP > NP, p=0.021).[11]

4. Individuals in the coordinated peer punishment and centralized punishment institutions make higher normative requests than members of the same institution under exogenous assignment in the early periods. Figure A.10 displays the cumulative density functions for individual normative requests in both institutions under endogenous selection and exogenous assignment for (a) the first period and (b) the first five periods. In all cases, the normative request under endogenous selection

[11]See Section A.3.4 for lack of evidence for bias in completing Part 3. For those who completed Part 3, 30 subjects selected NP in the first period of Part 2, 33 selected PP, and 19 selected CP. Given the similar baseline contributions (Figure A.9a) and contributions in the initial period of Part 2 (Figure A.5a), we find it reasonable to pool the SVO scores across the two punishment institutions for the statistical test.



(a) Period 1.



(b) Periods 1-5.

Figure A.10: Cumulative density functions of normative requests in (a) the first period and (b) the first five periods. In all cases, the normative request under endogenous selection first-order stochastically dominates the normative request under exogenous assignment. Density functions are based on individual normative requests in each period.

first-order stochastically dominates the normative request under exogenous assignment. In the first five periods of the endogenous selection treatment more than 80% of the subjects in PP had a normative request of 18 or more while only about 30% of the subjects had similarly high requests under exogenous assignment. Likewise, almost 90% of subjects in the CP institution requested a contribution level of 17 or more when they self-selected into this institution while only 60% had similarly high requests under exogenous assignment. These results show that the vast majority

of subjects requested very high contribution levels right from the beginning under endogenous selection, but a substantial number of subjects were satisfied with lower requests under exogenous assignment.

Taken together, these results suggest that subjects adopting punishment institutions under endogenous selection quickly established a cooperative culture. These institutions attracted a high share of prosocial individuals who quickly established high normative requests; these requests successfully coordinated the whole group to high contribution levels such that little punishment was necessary to enforce the widely agreed high contribution norm. In contrast, exogenous assignment of subjects to institutions puts sand into the gears of cooperation. It prevents the self-selection of prosocial individuals and causes substantial adjustment costs during the initial phases because subjects demand lower contributions and cooperate less, which then requires higher punishment costs to establish cooperation.

Why can't the subjects under exogenous assignment coordinate on high normative requests in the early periods? One reason could be that subjects in the exogenous assignment treatments are aware that the institution members are not very cooperative; therefore, trust must be built up over time and incrementally from lower initial requests. While this question is interesting, we cannot address it with our data; thus, it remains open for future investigation.

One final result emerged from our data that we did not anticipate in advance. Our centralized punishment institution was motivated by the presence of such institutions across societies from small-scale tribes to international organizations such as the United Nations Security Council, and we anticipated that centralization would lower the overall amount of punishment by removing the problem of coordinating punishment without communication under peer punishment. Antisocial punishment, in which above-average contributors are punished, is commonly observed under peer punishment (Herrmann et al., 2008). We find that centralization has an important effect on antisocial punishment.

Result A.5. *Centralization eliminates antisocial punishment.*

	Endogenous Selection		Exogenous Assignment	
	Peer Punishment	Central Punishment	Peer Punishment	Central Punishment
Own negative deviation from average contribution	0.234 (0.000)***	0.080 (0.001)***	0.367 (0.000)***	0.195 (0.001)***
Own positive deviation from average contribution	0.212 (0.000)***	-0.097 (0.022)**	0.146 (0.029)**	-0.410 (0.003)***
N	1,008	1,323	1,008	1,323
Left-censored	886	1,194	825	1,167
Uncensored	122	129	183	156
Number of clusters	11	11	11	11
Log-likelihood	-582.82	-589.61	-778.60	-821.82
AIC	1171.64	1185.22	1563.20	1649.64
ΔAIC	79.38	26.04	96.24	41.22
Bootstrap samples	9,999	9,999	9,999	9,999

Table A.4: Punishment received based on deviation from average group contribution during the period. Tobit regressions with robust standard errors clustered by matching group. Coefficients reported are average partial effects. Bootstrapped p-values given in parentheses based on 9,999 bootstrap samples computed using pairs cluster bootstrap-t with clustered standard errors. ΔAIC is the difference in Akaike Information Criterion for the reported regression compared to a model with deviations from normative request as independent variables; values exceeding 10 are considered very strong evidence in support of the regressions presented in the table.

***Significant at the 1 percent level.

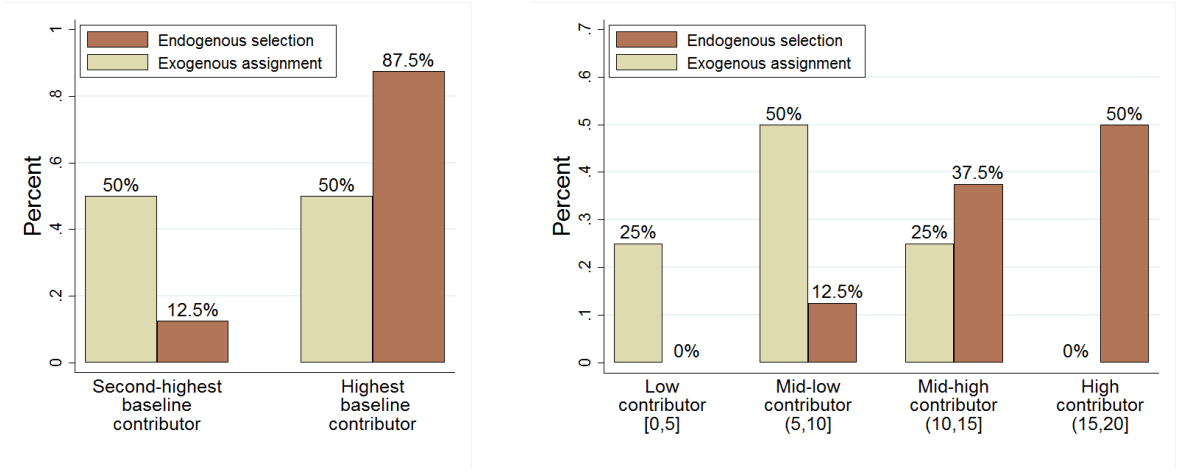
**Significant at the 5 percent level.

*Significant at the 10 percent level.

Evidence for Result 5. Table A.4 contains average partial effects from Tobit regressions estimating the effects of both negative and positive deviations from the average contribution. In all cases, larger negative deviations from the average are met with significantly higher amounts of punishment. Antisocial punishment remains a problem in coordinated peer punishment, as larger *positive* deviations are met with significantly higher amounts of punishment (average partial effect is 0.212 under endogenous selection and 0.146 under exogenous assignment). However, the opposite is found in centralized punishment, where larger positive deviations are met with significantly lower amounts of punishment (average partial effect is -0.097 under endogenous selection and -0.410 under exogenous assignment). We also considered the possibility that deviations from the normative re-

quest were the basis for punishment, but this model performed substantially worse than deviations from the average contribution. We report differences in Akaike Information Criterion, ΔAIC, in Table A.4; these values ranged from 26.04 to 79.38 in our data, and $\Delta AIC > 10$ is generally considered very strong evidence in support of the favored model (Burnham and Anderson, 2002).

Why should centralization eliminate antisocial punishment? Some recent papers do not find the same result (Fischer et al., 2013; Grechenig et al., 2013). The critical difference appears to be that we allow institution members to select the central authority, while these other recent papers exogenously assign the role randomly and keep it fixed throughout the experiment. In particular, it is often speculated that antisocial punishment comes from below-average contributors targeting above-average contributors either to send a message not to punish low contributions or to retaliate against received punishment (Herrmann et al., 2008). In our experiments, institution members tended to delegate authority to high average contributors, suggesting that they chose the "right" people and did not elect antisocial punishers. Figure A.11 illustrates delegation to high contributors in the first period of Part 2. Authority was always delegated to one of the two highest baseline contributors (Figure A.11a); the highest contributor was elected 87.5% of the time under endogenous selection but only 50% of the time under exogenous assignment. Figure A.11b illustrates that these contribution types, however, are very different between treatments. Under endogenous selection, 50% of the initial authorities contributed at least 16 points to the group project on average during Part 1 and 87.5% contributed at least 11 points on average. In contrast, 75% of the initial authorities in the exogenous assignment treatment contributed half of their endowment (10 points) or less on average. Thus, while subjects tend to delegate authority to the highest contributors in a relative sense (Figure A.11a), there is a difference in absolute contributions across treatments which is due to the self-selection of prosocial individuals in the endogenous selection treatment.



(a) Contribution rank

(b) Contribution type

Figure A.11: Baseline contribution behavior (Part 1) for subjects elected as central authority in first period of Part 2 (period 6 overall).

A.5 Discussion and concluding remarks

Institutions "are the humanly devised constraints that shape human interaction" (North, 1990). People routinely make decisions that determine who they interact with and in which institutions those interactions occur. As a result of these endogenous selection effects and varying institutions, real-world interactions may differ substantially from behavior in typical lab experiments in which interaction partners are random, institutions are exogenously imposed, or – most commonly – both.

There has been a recent surge in the economics literature exploring the role of endogenous group formation on cooperation and a corresponding surge exploring the effect of endogenous institution formation within fixed groups. By only varying one of these two aspects, the other factor, which commonly occurs outside the lab and which may be critical to sustain cooperation, is overlooked; while this approach is useful for identification of a single effect, it also prevents a fuller understanding of cooperation outside the lab. Moreover, the institutional options are often very limited and may not reflect institutions that are adopted in natural environments.

Our goal in this paper is to provide subjects with the opportunity to select into institutional environments in a public goods game and only interact with others who also

select into the same institution. Selection occurs every period, allowing us to observe migration patterns and changes in behavior across time. We allow subjects to select into four institutions. The first two institutions, no punishment and costly peer punishment, are commonly used in lab experiments. In the third institution, we add a cheap-talk normative request to peer punishment. In the fourth, we maintain the cheap-talk normative request but now allow the institution members to elect one member to assign all punishment that period but socialize the cost so that the burden is shared equally by all institution members.

Our first major result provides new evidence on which institutions people will adopt when given the option. Subjects are initially evenly divided among the institutions with no punishment, peer punishment with a normative request, and central punishment with a normative request; the peer punishment institution (without a normative request) used in many studies is almost never adopted. This result questions the inferences we can make about behavior outside the lab based on most previous experiments, as these experiments have typically relied on institutions that subjects would not normally adopt.

We also find several other interesting results. We demonstrate that prosocial individuals are more likely to select into punishment institutions in early periods. We also find that the cheap-talk normative request in these punishment institutions is used as a coordination device for contributions but not for determining punishment, which appears to be driven by deviations from the average contribution. The central punishment institution is able to eliminate the initial inefficiency that is usually observed with peer punishment and is also able to eliminate antisocial punishment.

Our work demonstrates the ability to capture several complex phenomena related to cooperation in an experimental design that allows these phenomena to be disentangled in the lab. The results shed new light on the early stages of institutional emergence and cooperation and how they coevolve in a manner that describes many real-world settings. Our design leaves open the possibility of many extensions that incorporate realistic extensions of the coevolutionary process, of which we suggest a few obvious

candidates: costly migration and entry restrictions; permanent formal authorities who can extract rents; increasing group size and imperfect information; and the process of internalizing norms as the basis for punishment. These extensions provide fertile ground for new experiments and allow for a better understanding of the coevolution of cooperative behavior and institutions in natural environments.

A.6 References

Acemoglu, D., D. Cantoni, S. Johnson, and J. A. Robinson (2011): "The Consequences of Radical Reform: The French Revolution," *American Economic Review*, 101, 3286–3307.

Ahn, T. K., R. M. Isaac, and T. C. Salmon (2008): "Endogenous group formation," *Journal of Public Economic Theory*, 10, 171–194.

——— (2009): "Coming and going: Experiments on endogenous group sizes for excludable public goods," *Journal of Public Economics*, 93, 336–351.

Andreoni, J. and J. Miller (2002): "Giving according to GARP: An experimental test of the consistency of preferences for altruism," *Econometrica*, 70, 737–753.

——— (2003): "The algebra of assortative encounters and the evolution of cooperation," *International Game Theory Review*, 5, 211–228.

Berkowitz, D., K. Pistor, and J.-F. Richard (2003a): "Economic development, legality, and the transplant effect," *European Economic Review*, 47, 165–195.

——— (2003b): "The transplant effect," *American Journal of Comparative Law*, 51, 163–203.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates?" *Quarterly Journal of Economics*, 119, 249–275.

Bochet, O., T. Page, and L. Putterman (2006): "Communication and punishment in voluntary contribution experiments," *Journal of Economic Behavior and Organization*, 60, 11–26.

Bochet, O. and L. Putterman (2009): "Not just babble: Opening the black box of communication in a voluntary contribution experiment," *European Economic Review*, 53, 309–326.

Boehm, C., C. Antweiler, I. Eibl-Eibesfeldt, S. Kent, B. M. Knauft, S. Mithen, P. J. Richerson, and D. S. Wilson (1996): "Emergency Decisions, Cultural-Selection Mechanics, and Group Selection [and Comments and Reply]," *Current Anthropology*, 37, 763–793.

Bolton, G. E. and A. Ockenfels (2000): "ERC: A theory of equity, reciprocity, and competition," *American Economic Review*, 166–193.

Burnham, K. P. and D. R. Anderson (2002): *Model selection and multi-model inference: a practical information-theoretic approach*, Springer.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 90, 414–427.

Charness, G. and M. Rabin (2002): "Understanding social preferences with simple tests," *Quarterly Journal of Economics*, 117, 817–869.

Dreber, A., D. G. Rand, D. Fudenberg, and M. A. Nowak (2008): "Winners don't punish," *Nature*, 452, 348–351.

Eshel, I. and L. L. Cavalli-Sforza (1982): "Assortment of encounters and evolution of cooperativeness," *Proceedings of the National Academy of Sciences*, 79, 1331.

Farrell, J. and M. Rabin (1996): "Cheap talk," *Journal of Economic Perspectives*, 10, 103–118.

Fehr, E. and U. Fischbacher (2003): "The nature of human altruism," *Nature*, 425, 785–791.

Fehr, E. and S. Gächter (2000): "Cooperation and punishment in public goods experiments," *American Economic Review*, 90, 980–994.

Fehr, E. and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 114, 817–868.

Fischbacher, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10, 171–178.

Fischbacher, U., S. Gächter, and E. Fehr (2001): "Are people conditionally cooperative? Evidence from a public goods experiment," *Economics Letters*, 71, 397–404.

Fischer, S., K. R. Grechenig, and N. Meier (2013): "Cooperation under Punishment: Imperfect Information Destroys it and Centralizing Punishment Does Not Help," *MPI Collective Goods Preprint*.

Forsythe, R., J. L. Horowitz, N. E. Savin, and M. Sefton (1994): "Fairness in simple bargaining experiments," *Games and Economic behavior*, 6, 347–369.

Gächter, S., E. Renner, and M. Sefton (2008): "The long-run benefits of punishment," *Science*, 322, 1510–1510.

Grechenig, K., A. Nicklisch, and C. Thöni (2013): "Information-sensitive Leviathans – the emergence of centralized punishment," Unpublished working paper.

Greiner, B. (2004): "An Online Recruitment System for Economic Experiments," in *Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht 63, Göttingen: Ges. für Wiss. Datenverarbeitung*, ed. by K. Kremer and V. Macho, 79–93.

Gürerk, Ö., B. Irlenbusch, and B. Rockenbach (2006): "The competitive advantage of sanctioning institutions," *Science*, 312, 108–111.

——— (2011): "Voting with feet: community choice in social dilemmas," Unpublished working paper.

——— (2013): "On cooperation in open communities," Unpublished working paper.

Hamman, J. R., R. A. Weber, and J. Woon (2011): "An experimental investigation of electoral delegation and the provision of public goods," *American Journal of Political Science*, 55, 738–752.

Henrich, J. (2004): "Cultural group selection, coevolutionary processes and large-scale cooperation," *Journal of Economic Behavior and Organization*, 53, 3–35.

Henrich, J., J. Ensminger, R. McElreath, A. Barr, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, et al. (2010): "Markets, religion, community size, and the evolution of fairness and punishment," *Science*, 327, 1480–1484.

Henrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, et al. (2006): "Costly punishment across human societies," *Science*, 312, 1767–1770.

Herrmann, B., C. Thöni, and S. Gächter (2008): "Antisocial punishment across societies," *Science*, 319, 1362–1367.

Isaac, R. M. and J. M. Walker (1988a): "Communication and free-riding behavior: The voluntary contribution mechanism," *Economic Inquiry*, 26, 585–608.

Kaplan, H., M. Gurven, K. Hill, and A. M. Hurtado (2005): "The natural history of human food sharing and cooperation: a review and a new multi-individual approach to the negotiation of norms," in *Moral sentiments and material interests: The foundations of cooperation in economic life*, ed. by H. Gintis, S. Bowles, R. Boyd, and E. Fehr, 75–113.

Kimbrough, E. O., V. L. Smith, and B. J. Wilson (2008): "Historical Property Rights, Sociality, and the Emergence of Impersonal Exchange in Long-Distance Trade," *American Economic Review*, 98, 1009–1039.

Kosfeld, M., A. Okada, and A. Riedl (2009): "Institution formation in public goods games," *American Economic Review*, 1335–1355.

Marlowe, F. W., J. C. Berbesque, A. Barr, C. Barrett, A. Bolyanatz, J. C. Cardenas, J. Ensminger, M. Gurven, E. Gwako, J. Henrich, et al. (2008): "More 'altruistic' punishment in larger societies," *Proceedings of the Royal Society B: Biological Sciences*, 275, 587–592.

Mathew, S. and R. Boyd (2011): "Punishment sustains large-scale cooperation in prestate warfare," *Proceedings of the National Academy of Sciences*, 108, 11375–11380.

Murphy, R. O., K. A. Ackermann, and M. J. Handgraaf (2011): "Measuring Social Value Orientation," *Judgment and Decision Making*, 6, 771–781.

North, D. C. (1990): *Institutions, institutional change and economic performance*, Cambridge university press.

Nowak, M. A. (2006): "Five rules for the evolution of cooperation," *Science*, 314, 1560–1563.

——— (forthcoming): "Historical development," in *Handbook of Economic Growth, Volume 2*, ed. by P. Aghion and S. N. Durlauf, North-Holland.

Ostrom, E. (1990): *Governing the commons: The evolution of institutions for collective action*, Cambridge University Press.

——— (1998): "A behavioral approach to the rational choice theory of collective action: Presidential address, American Political Science Association, 1997," *American Political Science Review*, 1–22.

Ostrom, E., J. Walker, and R. Gardner (1992): "Covenants with and without a sword: Self-governance is possible," *American Political Science Review*, 404–417.

Pennisi, E. (2005): "How did cooperative behavior evolve?" *Science*, 309, 93–93.

Sally, D. (1995): "Conversation and Cooperation in Social Dilemmas A Meta-Analysis of Experiments from 1958 to 1992," *Rationality and society*, 7, 58–92.

Sutter, M., S. Haigner, and M. G. Kocher (2010): "Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations," *Review of Economic Studies*, 77, 1540–1566.

Tiebout, C. M. (1956): "A pure theory of local expenditures," *Journal of Political Economy*, 64, 416–424.

Wiessner, P. (2005): "Norm enforcement among the Ju/'hoansi Bushmen," *Human Nature*, 16, 115–145.

Wilson, R. K. and J. Sell (1997): ""Liar, Liar..." Cheap Talk and Reputation in Repeated Public Goods Settings," *Journal of Conflict Resolution*, 41, 695–717.

A.7 Appendix

A.7.1 Cooperative equilibria in Coordinated Central Punishment

Here, we characterize one set of cooperative equilibria in a one-shot version of the public goods game with centralized punishment and cheap-talk normative request. We use the inequity aversion model of Fehr and Schmidt (1999) for simplicity and tractability. One could alternatively use other popular inequity aversion models such as Bolton and Ockenfels (2000) or Charness and Rabin (2002). With Charness and Rabin (2002) preferences, one needs to include their *demerit profile* to capture reciprocity; the simpler version with only the disinterested social-welfare criterion cannot explain punishment behavior, as punishment strictly decreases own payoff and social surplus while weakly decreasing the minimum payoff in the institution (see Appendix 1 in their paper for both versions of the model).

Utility under inequity aversion due to Fehr and Schmidt (1999)

Let $\pi = (\pi_1, \ldots, \pi_n)$ denote the vector of monetary payoffs. Then, player i's utility is given by

$$u_i(\pi) = \pi_i - \alpha_i \left(\frac{1}{n-1}\right) \sum_{j \neq i} \Big[\max\{x_j - x_i, 0\}\Big] - \beta_i \left(\frac{1}{n-1}\right) \sum_{j \neq i} \Big[\max\{x_i - x_j, 0\}\Big] \quad \text{(A.7)}$$

with the conditions that $\beta_i \leq \alpha_i$ and $0 \leq \beta_i \leq 1$. In the utility function, α_i captures the decrease in utility due to disadvantageous inequality, while β_i captures the decrease in utility due to advantageous inequality.

Proposition 1. *Without loss of generality, order the values of* α_i *such that* $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_n$. *Suppose there exists* $q \in \{1, 2, \ldots, n\}$ *such that preferences satisfy* $(m/n) + \beta_j \geq$

1 *and*

$$c < nr\alpha_j \tag{A.8}$$

for all $j \geq q$ *and that* $\alpha_j = \beta_j = 0$ *for the remaining players, where* $j < q$*. In the public goods game with centralized punishment, all strategy profiles satisfying the following three properties constitute subgame perfect equilibria:*

1. *Voting results in one of the players* $j \geq q$ *being elected as central authority with certainty. (Note that no specific member needs to be elected; moreover, ties are acceptable as long as* $j \geq q$ *for all candidates tied for the largest number of votes.)*

2. *During the contribution stage,* $g_i = g \in [0, e_1]$ *for all* i*.*

3. *If one of the players contributes* $g_i < g$*, the central authority assigns* $(g - g_i)/r$ *punishment points to player* i*.*

4. *Off the equilibrium path, any player* $j < q$ *elected to be the central authority will not punish because* $\alpha_j = \beta_j = 0$*. Using backwards induction, this lack of punishment results in zero contributions at the contribution stage.*

Proof. Suppose that one of the players contributes $g_i < g$ during the contribution stage. Let P denote the punishment assigned by the central authority to player i. Monetary payoffs are

$$\pi_i = y - g_i + \left(\frac{m}{n}\right)[(n-1)g + g_i] - rP - \left(\frac{c}{n}\right)P \tag{A.9}$$

for player i and

$$\pi_j = y - g_+ \left(\frac{m}{n}\right)[(n-1)g + g_i] - \left(\frac{c}{n}\right)P \tag{A.10}$$

for all $j \neq i$.

Notice, importantly, that player i's monetary payoff is reduced by both the received punishment and by her share of the cost of punishment, which is shared equally by all group members. We propose that the value of P that equalizes final payoffs will constitute

an equilibrium.

$$\begin{aligned}
\pi_i &= \pi_j \\
y - g_i + \left(\frac{m}{n}\right)[(n-1)g + g_i] - rP - \left(\frac{c}{n}\right)P &= y - g_+ \left(\frac{m}{n}\right)[(n-1)g + g_i] - \left(\frac{c}{n}\right)P \\
-g_i - rP - \left(\frac{c}{n}\right)P &= -g - \left(\frac{c}{n}\right)P \\
g - g_i &= rP \\
P &= \frac{g - g_i}{r}.
\end{aligned}$$

While it should be clear that π_i is less than the equilibrium payoff, we include the algebra here for completeness. The monetary payoff from equilibrium is $\pi = y + (m-1)g$.

$$\begin{aligned}
\pi = y + (m-1)g &> y - g_+ \left(\frac{m}{n}\right)[(n-1)g + g_i] \\
&> y - g_+ \left(\frac{m}{n}\right)[(n-1)g + g_i] - \left(\frac{c}{n}\right)P \\
&= \pi_j \\
&= \pi_i.
\end{aligned}$$

Therefore, player i has no incentive to deviate, conditional on the punishment threat being credible.

Suppose the central authority reduces P by ε. The central authority's monetary payoff increases by $(c/n)\varepsilon$, as does the payoff of all other institution members (including player i). Unlike peer punishment, there is no inequity between the central authority and the other full contributors, as the cost of punishment is shared equally by all institution members. Player i's monetary payoff increases by $r\varepsilon$ from the reduction in received punishment and by $(c/n)\varepsilon$ from the reduction in the cost of assigned punishment. Therefore, the central authority, player j, suffers a non-monetary reduction in utility due to disadvantageous inequality in the amount of $\alpha_j r\varepsilon$. Thus, if $\alpha_j r\varepsilon > (c/n)\varepsilon$, the punishment threat is credible and the central authority prefers not to deviate from the proposed equilibrium.

Since the ε term is common, the requirement reduces to $c < (nr\alpha_j)$, which is the condition stated in the proposition.

Notice also that the central authority does not have an incentive to punish player i beyond the point of equal payoffs. Any excess punishment causes a reduction in the central authority's utility by decreasing the monetary payoff due to increased costs of punishment and by increasing advantageous inequality with respect to player i.

Finally, we need to demonstrate that the central authority will not deviate in the contribution stage. The argument is the same here as in Fehr and Schmidt (1999)'s proof of their Proposition 5. The central authority can reduce her contribution to the public good by $\varepsilon > 0$ and increase her material payoff by $(1 - (m/n))\varepsilon$. Doing so creates advantageous inequality of ε relative to each of the other institution members, causing an overall decrease in utility of $\beta_i(1/(n-1))(n-1)\varepsilon = \beta_i\varepsilon$. The central authority will only deviate if $(1 - (m/n))\varepsilon > \beta_i\varepsilon$ or, equivalently, $(m/n) + \beta_i < 1$. This last condition is ruled out by assumption in the proposition. Therefore, the central authority will never deviate in the contribution stage.

The condition on voting is trivial.

□

A.7.2 Social Value Orientation

The Social Value Orientation (SVO) measure of Murphy et al. (2011) consists of six allocation decisions between oneself and one other anonymous individual. The SVO scale can be used to classify individuals as (1) altruistic, (2) prosocial, (3) individualistic, or (4) competitive.[12] The decision-making criteria for the four social preference types are given in Table A.5.

[12]The 6-item SVO scale cannot distinguish between prosocial individuals who are efficiency maximizers (Type 2.a) and prosocial individuals who are inequity averse (Type 2.b); an additional 9 items are provided by (Murphy et al., 2011) to distinguish between these two subtypes but are not relevant for the other classifications. Therefore, we rely on the 6-item measure as it provides a useful measure for all subjects. Moreover, distinguishing between efficiency-maximization and inequity aversion is not necessary or useful, as groups converge to the Pareto-dominant cooperative equilibrium of full contributions, which is consistent with both subtypes.

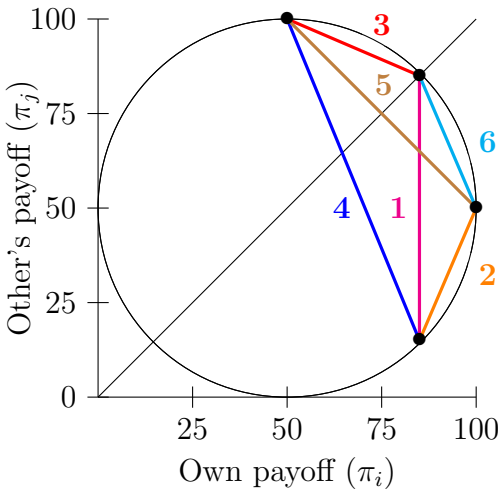


Figure A.12: Social Value Orientation (SVO) allocation decisions. Each numbered line represents the set of options for one decision. Options consist of nine equally spaced allocations along the line. *Source:* Adapted from Murphy et al. (2011).

Type		Condition	Intuition
1	Altruistic	$\max\{\pi_j\}$	Maximize other's payoff
2.a	Prosocial (efficiency)	$\max\{\pi_i + \pi_j\}$	Maximize total payoff
2.b	Prosocial (inequity averse)	$\min\{\pi_i - \pi_j\}$	Minimize relative payoff
3	Individualistic	$\max\{\pi_i\}$	Maximize own payoff
4	Competitive	$\max\{\pi_i - \pi_j\}$	Maximize relative payoff

Table A.5: Decision criteria for Social Value Orientation types.

The six allocation decisions are displayed by the numbered lines in Figure A.12.[13] The SVO score is measured in degrees and is given by

$$SVO_i = \arctan\left(\frac{\overline{\pi_j} - 50}{\overline{\pi_i} - 50}\right) \tag{A.11}$$

where $\overline{\pi_i}$ ($\overline{\pi_j}$) is the average amount allocated to oneself (other person) over the six allocation decisions. Higher SVO scores indicate stronger social preferences.

[13]For example, line 2 in Figure A.12 represents choosing among 9 equally spaced allocations ranging from (100,50) to (85,15), where the first entry is payoff to self (π_i) and the second entry is payoff to other (π_j). An individualistic person maximizes own payoff and would choose (100,50). A competitive person maximizes the relative difference between her own payoff and her counterpart's payoff; such a person would choose (85,15) because it yields the maximal relative payoff difference of 70.

A.7.3 Experimental instructions: Part 1

Instructions for Part I

Thank you for participating in today's study on decision-making. For participating in today's study, you will receive a fixed payment of 10 CHF in addition to earnings based on your decisions during the study. There are two parts to today's study. This is Part I. You will receive instructions for Part II of the study after completing Part I. During both parts of the study, you will earn points. At the end of the study, these points will be converted to Swiss Francs at the following rate:

1 point = 0.05 CHF

Your earnings from Part II will be added to your earnings from Part I. At the end of the study, all of your earnings and your fixed payment will be paid to you in cash.

If you have any questions during the study, please raise your hand and wait for someone to come to you. Please do not make any noise or try to communicate with other participants during the study. Also, please turn off any electronic devices such as mobile phones and iPods. If you violate these rules, you may be asked to leave the study without receiving any payment.

Each participant in the study will be assigned a participant number when Part I begins and will interact with the same set of people throughout Part I. This number is assigned randomly by a computer. You will see your participant number each period, and participant numbers will not change. For example, if you were Participant 5 in period 1, you will still be Participant 5 in period 2. In addition, Participant 4 in period 2 will be the same person in who was Participant 4 in period 1. Your group will have up to 12 members in total, including yourself.

This part of the study is divided into 10 periods. In each period, you receive 20 points. You will then decide how many points you want to contribute to a project and how many points to send to a private account.

Each point in the private account adds 1 point to your earnings. Earnings from the project are calculated in the following manner: the average number of points contributed to the project will be multiplied by 1.5 and added to your earnings. In addition, every other member of the group will receive the same amount from the project. Therefore, points contributed to the project add to both your earnings and the earnings of the other group members. Earnings for each group member are calculated in the same way. You will only be asked how many points you would like to contribute to the project. All remaining points are sent to your private account.

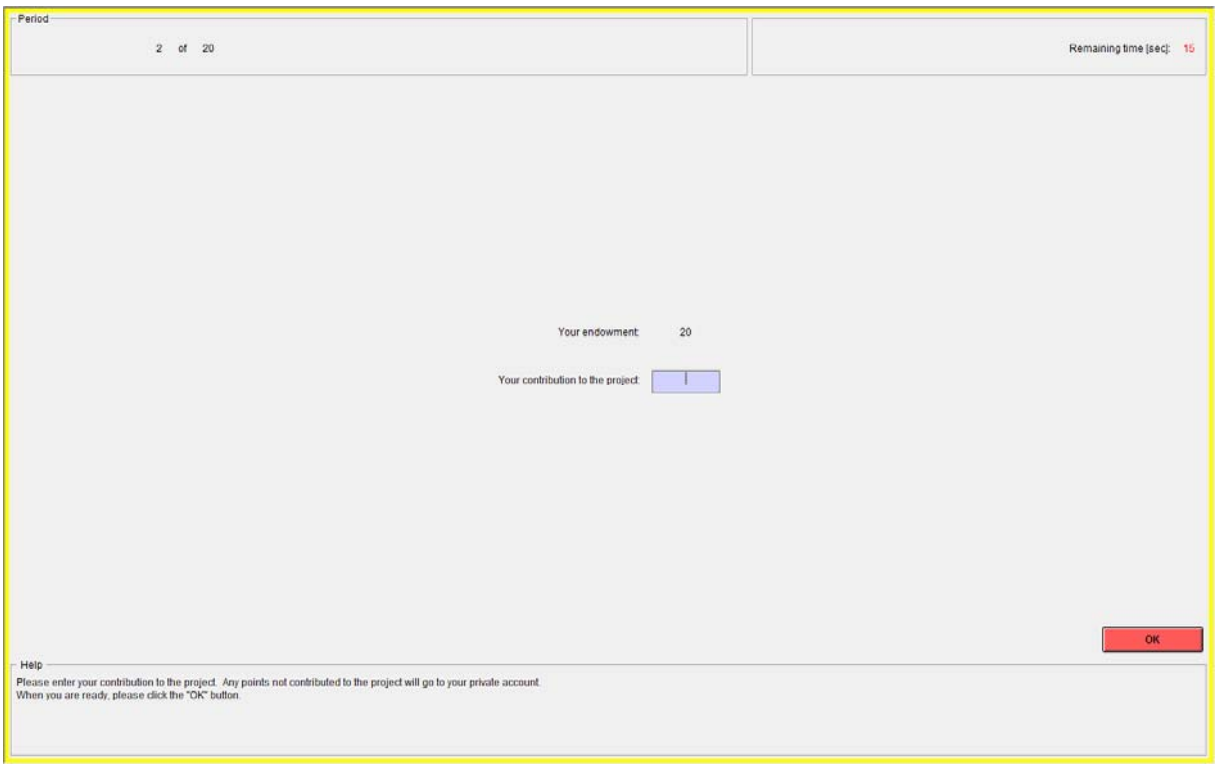
Earnings in each period are calculated in the following manner:

20 – (your contribution to the project) + 1.5*(average contribution to the project)

Your earnings in each period are added together to determine your final earnings from Part I.

Example. There are 2 group members. Participant 1 contributes 5 points to the project (and 15 points to the private account). Participant 2 contributes 15 points to the project (and 5 points to the private account). The average number of points contributed to the project is 10 points ([5 + 15]/2 = 20/2 = 10). Each group member receives 15 points (10*1.5 = 15) from the project. In addition, Participant 1 earns 15 points from the private account for a total of 30 points (15 + 15 = 30). Participant 2 earns 5 points from the private account for a total of 20 points (15 + 5 = 20).

The decision screen will look like this:

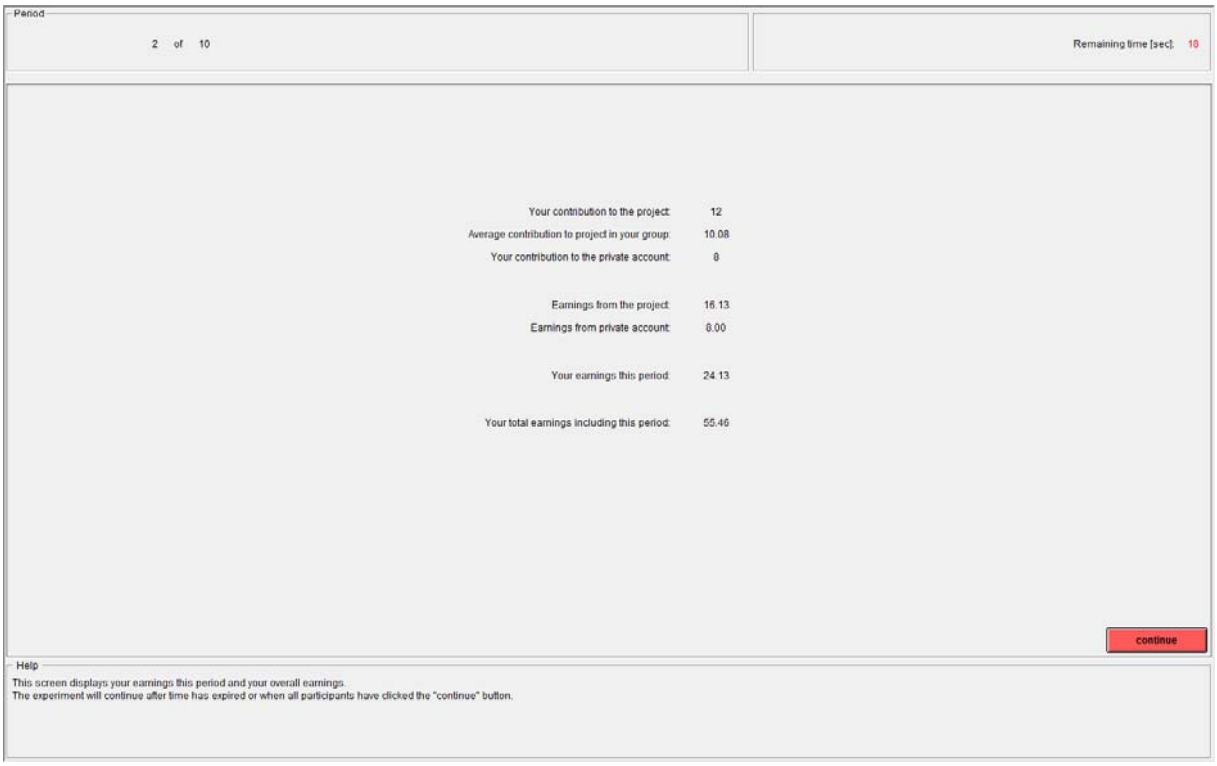


After all group members make their decisions, you will see two additional screens. The first screen will display each group member's participant number and contribution to the project.

The screen will look like this:

The second screen will show your decisions and earnings from this period and your overall earnings for Part I.

The screen will look like this:



Please complete the questions on the next sheet. If you have any questions or are finished answering the questions, please raise your hand. When everyone has completed the questions correctly, we will begin Part I of the study.

The following table may be helpful when completing the questions.

1.5*1 = 1.5	1.5*6 = 9	1.5*11 = 16.5	1.5*16 = 24
1.5*2 = 3	1.5*7 = 10.5	1.5*12 = 18	1.5*17 = 25.5
1.5*3 = 4.5	1.5*8 = 12	1.5*13 = 19.5	1.5*18 = 27
1.5*4= 6	1.5*9 = 13.5	1.5*14 = 21	1.5*19 = 28.5
1.5*5 = 7.5	1.5*10 = 15	1.5*15 = 22.5	1.5*20 = 30

A.7.4 Experimental instructions: Part 2

Instructions for Part II

As in Part I, you can earn points in Part II. At the end of the study, these points will be converted to Swiss Francs at the following rate:

1 point = 0.05 CHF

Your earnings from Part II will be added to your previous earnings from Part I and the fixed payment. At the end of the study, all of your earnings and your fixed payment will be paid to you in cash.

You will keep your participant number from Part I; if you were Participant 5 in Part I, you will still be Participant 5 in this part. You will also interact with the same set of 12 people from Part I; for example, Participant 4 in Part I will be the same person in Part II and will still be referred to as Participant 4.

This part of the study is divided into 20 periods. In each period, you will have the option to decide which group to join. Groups are defined by the rules of interacting with other group members. These rules will be described in more detail below. You can join a new group each period or stay in the same group. The four groups are Group A, Group B, Group C, and Group D.

Each period consists of group selection, which takes place at the very beginning of the period, followed by three stages.

- In **Group Selection**, you will decide which group to join.
- **Stage 1** will only take place in some groups. In Stage 1, if it takes place, the participants indicate how many points they think each group member should contribute to the project. In Stage 2, they will learn how many points the members of their group think each member should contribute. We call this the "group opinion."
- In **Stage 2**, which takes place in all four groups, you will receive 20 points and then make contribution decisions on the basis of the same rules as in Part I of the study.
- In **Stage 3**, you will receive another 20 points. In some groups, the members will be able to reduce the earnings of other group members by assigning deduction points to them.

In **Group Selection**, every participant is asked which group he would like to join. After the first period, this screen will also contain the average earnings in each group in the previous period and the number of people in each group in the previous period.

In **Stage 2**, the members in all groups will make contribution decisions according to the same rules as in Part I. You will receive 20 points at the beginning of Stage 2 and can send points to a private account and a project. Each point in the private account adds 1 point to your earnings. Earnings from the project are calculated in the following manner: The average number of points contributed to the project in your group will be multiplied by 1.5 and added to your earnings. In addition, each member

of the group will receive the same amount from the project. Therefore, points contributed to the project add to both your earnings and the earnings of the other group members. Earnings for each group member are calculated in the same way. You will only be asked how many points you would like to contribute to the project. All remaining points are sent to your private account.

Your earnings in Stage 2 are calculated in the following manner:

20 - (your contribution to the project) + 1.5 * (average contribution to the project)

The different groups that you can join will be explained briefly below. Afterwards, you will see the screens presented during one period. These screens will differ based on the group. You will also receive additional information about your group members' previous contributions and their contributions to the project during the current period. This information will be clear from the screen shots.

The Groups

In total there are four different groups that have partly different and partly identical rules. In Group Selection of a given period, the participants decide which group they want to join.

The **groups differ with regard to Stage 1**. Stage 1 only takes place in Groups C and D, but not in Groups A and B. In Stage 1, if it takes place, the participants indicate how many points they think each member of their group should contribute to the project.

The rules of all groups are identical with regard to Stage 2 of a given period. Recall that, in Stage 2, the group members can make a contribution to a project, i.e., they make contribution decisions on the basis of the same rules as in Part I.

In addition, the groups differ with regard to Stage 3. In Groups B and C, all group members can reduce other group members' earnings by assigning deduction points. In Group D, the group will elect one group member to assign all deduction points for the group. In Group A, no deduction points can be assigned.

	Stage 1 Formation of group opinion	Stage 2 Contribution to project	Stage 3 Assignment of deduction points	Who can assign deduction points in Stage 3?
Group A	-	✔	-	Nobody
Group B	-	✔	✔	All group members
Group C	✔	✔	✔	All group members
Group D	✔	✔	✔	Only the group member elected by the group

Group A (no formation of group opinion; no assignment of deduction points)

The rules for Group A are the same as the rules in Part I of the study, except for the 20 points that you now receive in Stage 3.

- **Stage 1** does not take place, i.e., the participants cannot indicate how many points each member of their group should contribute to the project.

- Before Stage 2, the members of Group A are informed about each group member's contribution to the project in the previous period.

- In **Stage 2**, members of Group A receive 20 points and make contribution decisions on the basis of the same rules as in Part I of the study.
  Exception: If you are the only person to join Group A in a given period, all 20 points will be sent directly to your private account.

- In **Stage 3**, each member of Group A receives an additional 20 points which are added directly to his earnings.

Group B (no formation of group opinion; every member assigns deduction points)

- **Stage 1** does not take place, i.e., the participants cannot indicate how many points each member of their group should contribute to the project.

- Before Stage 2, as in Group A, the members of Group B are informed about each group member's contribution to the project in the previous period.

- In **Stage 2**, members of Group B receive 20 points and make contribution decisions on the basis of the same rules as in Part I of the study.

Exception: If you are the only person to join Group B in a given period, all 20 points will be sent directly to your private account.

- In **Stage 3**, each member of Group B receives an additional 20 points which can be used to reduce the earnings of other group members. Members of Group B can reduce the earnings of others in the group by assigning deduction points. This decision is made after each group member has been informed about the other group members' contributions to the project. Assigning 1 deduction point to a group member will reduce that member's earnings by 3 points. Each deduction point you assign will reduce your earnings by 1 point. You can assign as many deduction points as you like to each group member, and you may assign deduction points to as many group members as you like. The only condition is that you cannot assign more than 20 deduction points in total. Any points remaining after you have assigned deduction points are added to your earnings.

Exception. Your earnings cannot be negative in any period. If your earnings are negative after assigning and receiving deduction points, you will receive 0 points at the end of the period.

Group C (formation of group opinion; every member assigns deduction points)

Group C is the same as Group B, but with one addition. In Stage 1, each member of Group C will decide how many points he thinks each group member should contribute to the project. Thus, after at least one participant has joined Group C, periods are structured as follows:

- In **Stage 1**, group members will answer the following question: **"How many points do you think each group member should contribute to the project?"** In Stage 2, the average number of points entered by the group members will be displayed on the screen when group members are asked to make a contribution decision.

- Before Stage 2, as in Groups A and B, the members of Group C are informed about each group member's contribution to the project in the previous period.

- **Stage 2** is the same in all groups; members of Group C receive 20 points and make contribution decisions on the basis of the same rules as in Part I of the study.
  Exception: If you are the only person to join Group C in a given period, all 20 points will be sent directly to your private account.

- **Stage 3** is the same as in Group B. Each group member receives 20 points and can assign deduction points to the other group members after he has been informed about the other group members' contributions to the project.

Exception. Your earnings cannot be negative in any period. If your earnings are negative after assigning and receiving deduction points, you will receive 0 points at the end of the period.

Group D (formation of group opinion; elected member assigns deduction points)

Group D has the same rules as Group C, but instead of all group members being in a position to assign deduction points to each other, the group can elect one member who will assign all the deduction points in Stage 3. Thus, periods are structured as follows:

- In **Stage 1**, group members will answer the following question: **"How many points do you think each group member should contribute to the project?"** In Stage 2, the average number of points entered by the group members will be displayed on the screen when group members are asked to make a contribution decision.

- In addition, before Stage 2, the **members of Group D can elect one member of the group who will then assign all the deduction points in Stage 3**. Before the participants elect this member, they are informed about each group member's contribution to the project in the previous period. In addition, they are also informed about who was selected to assign deduction points in the previous period. This participant may or may not be in Group D in the current period, depending on whether he left the group. The group member who receives the most votes will assign the deduction points. In case of a tie, one person will be randomly selected from those who received the most votes (with each one equally likely to be selected).

- **Stage 2** is the same as in all groups; members of Group D receive 20 points and make contribution decisions on the basis of the same rules as in Part I of the study.
  Exception: If you are the only person to join Group D in a given period, all 20 points will be sent directly to your private account.

- In **Stage 3** the elected member will assign deduction points to the other members of the group after he has observed the contributions of all group members. Each received deduction point reduces your earnings by 3 points. **However, in contrast to Groups B and C, the cost of the deduction points assigned will be shared equally by the group members**. For example, if there are 10 group members and 1 deduction point is assigned, then each group member's earnings will be reduced by 0.1 points (1/10 = 0.1). **The total number of deduction points that the elected group member can assign is given by 20*(number of group members)**. For example, if there are 3 group members, then the elected group member can assign up to 60 (20*3 = 60) deduction points in total.

Exception. Your earnings cannot be negative in any period. If your earnings are negative after assigning and receiving deduction points, you will receive 0 points at the end of the period.

Earnings

Final earnings in each period are calculated as follows:

(Earnings from Stage 2) + 20 - (3* received deduction points) - (cost of assigned deduction points)

If this amount is negative, then final earnings for the period are 0 points.

In Groups B and C,

cost of assigned deduction points = 1*(number of deduction points you assigned)

In Group D,

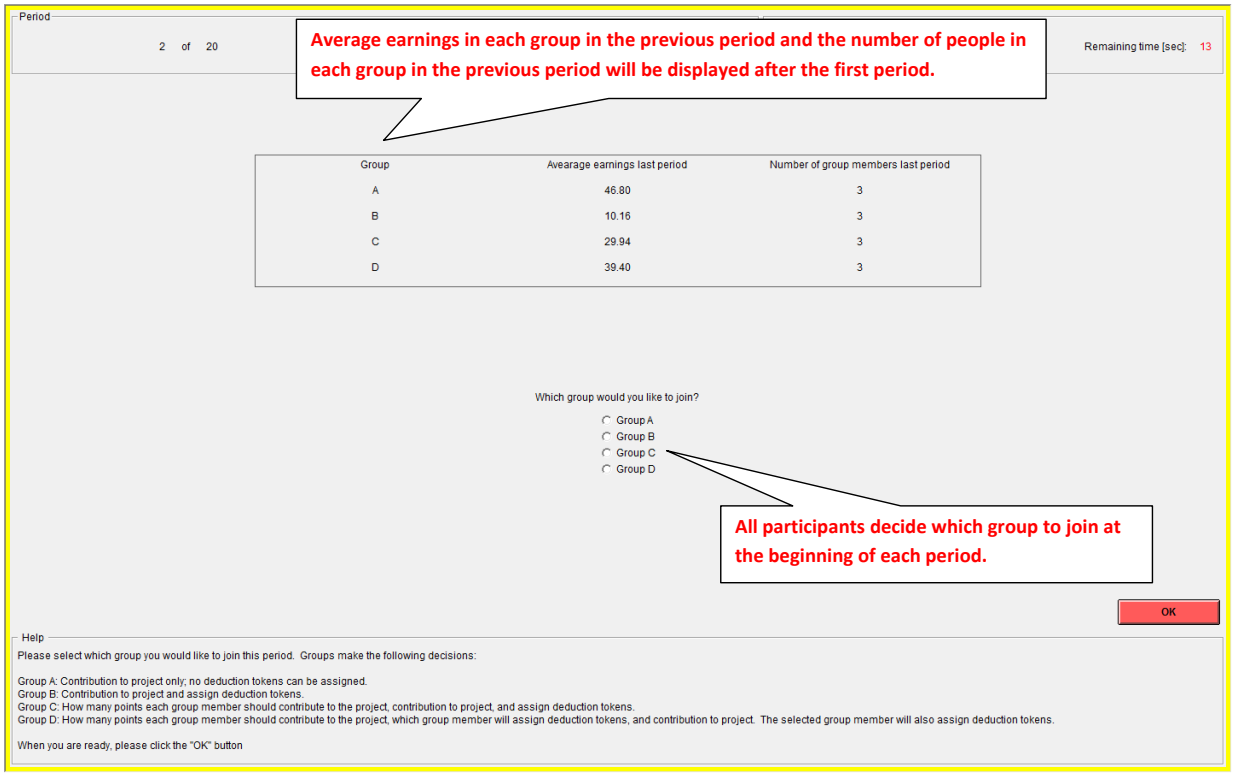
$$\textit{cost of assigned deduction points} = \left(\frac{\textit{total number of deduction points assigned by elected group member}}{\textit{number of group members}} \right)$$

Your final earnings from Part II are the sum of your earnings from each period. These earnings will be added to your earnings from Part I and your fixed payment.

On the following pages, you will see the screens shown during one period. These screens will differ based on the group. Afterwards, you will be asked to complete a few questions. If you have any questions or are finished answering the questions, please raise your hand. When everyone has completed the questions correctly, we will begin Part II of the study.

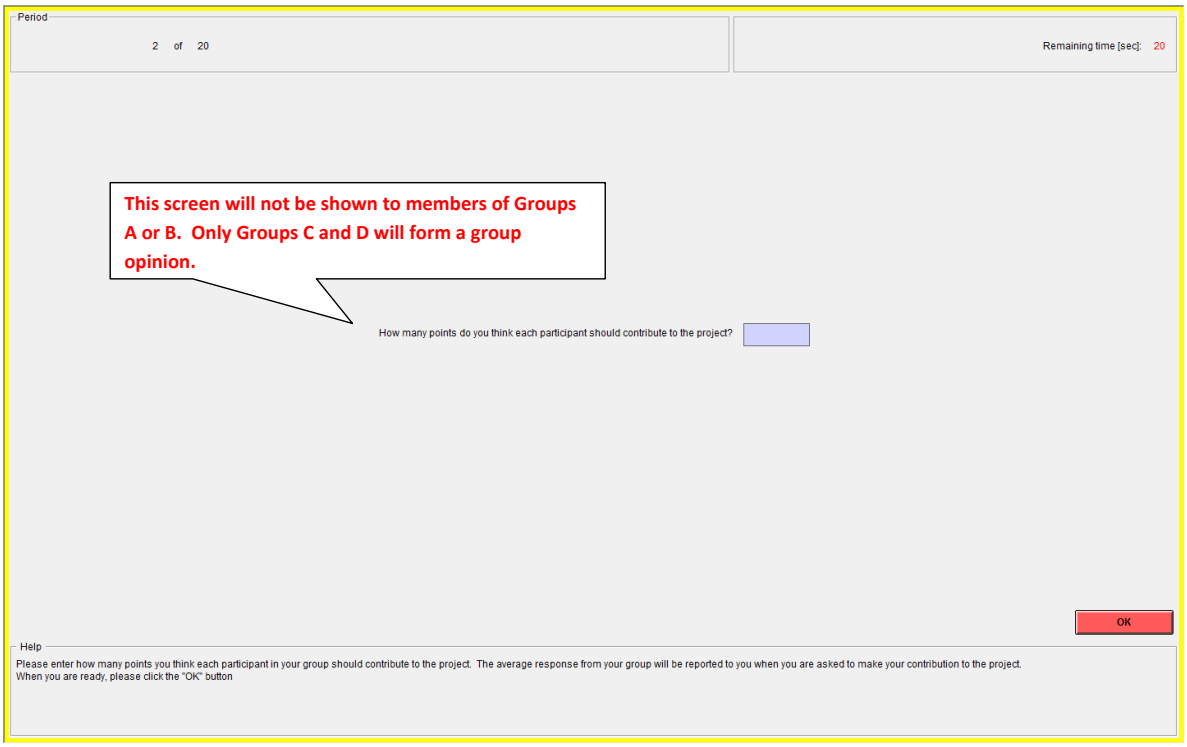
Group Selection

At the beginning of each period, all participants will see the **Group Selection** screen on which each participant is asked which group he would like to join. After the first period, this screen will also contain the average earnings in each group in the previous period and the number of people in each group in the previous period.



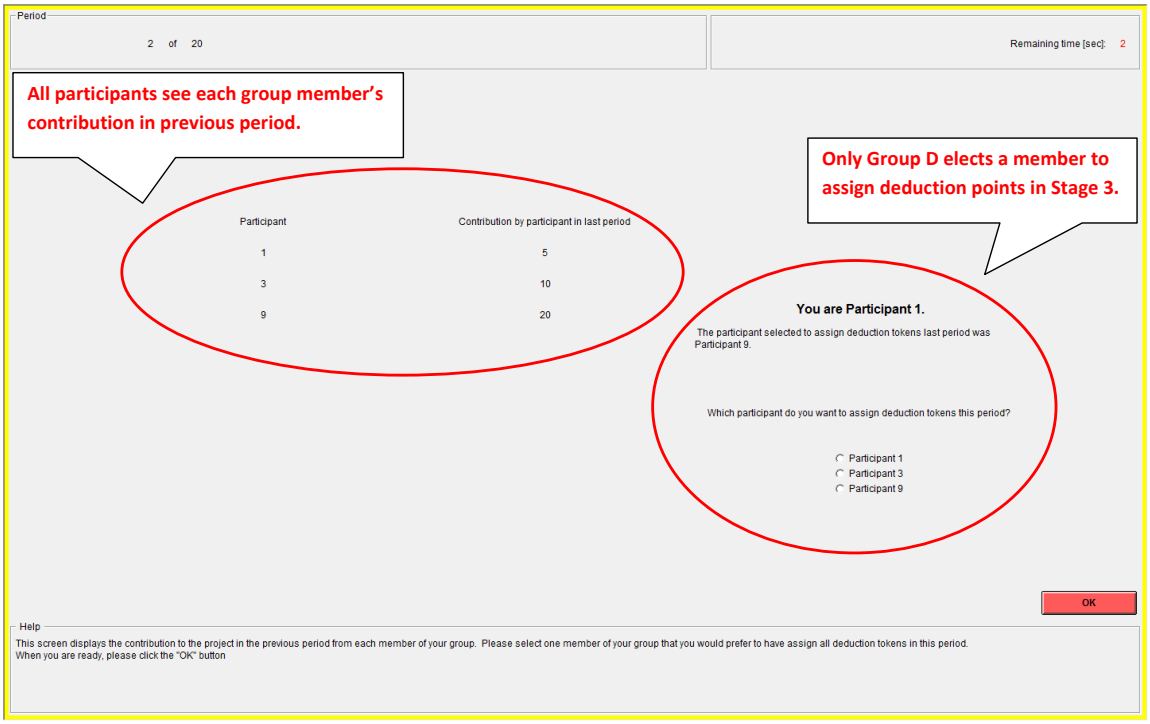
Stage 1: Group Opinion

In **Stage 1**, members of Group C and Group D will be asked to answer the following question: **"How many points do you think each group member should contribute to the project?"** In Stage 2, the average number of points entered by the group members will be displayed on the screen when group members are asked to make a contribution decision.



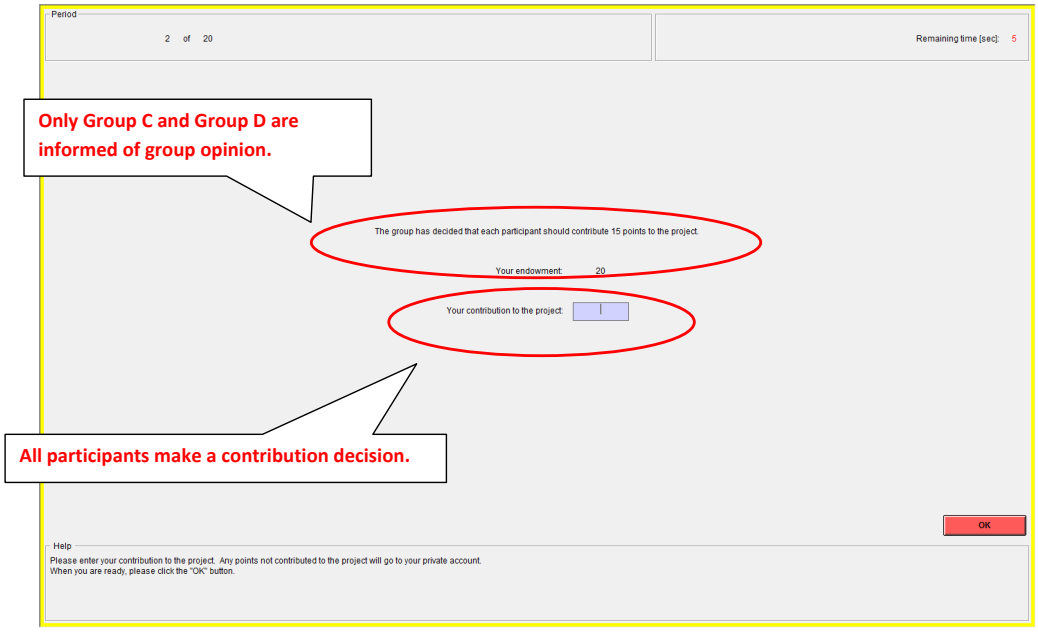
Before Stage 2: Previous contributions and election

Before Stage 2, all participants are informed of every group member's contribution to the project in the previous period. In addition, the members of Group D can elect one member of the group who will then assign all the deduction points in Stage 3.



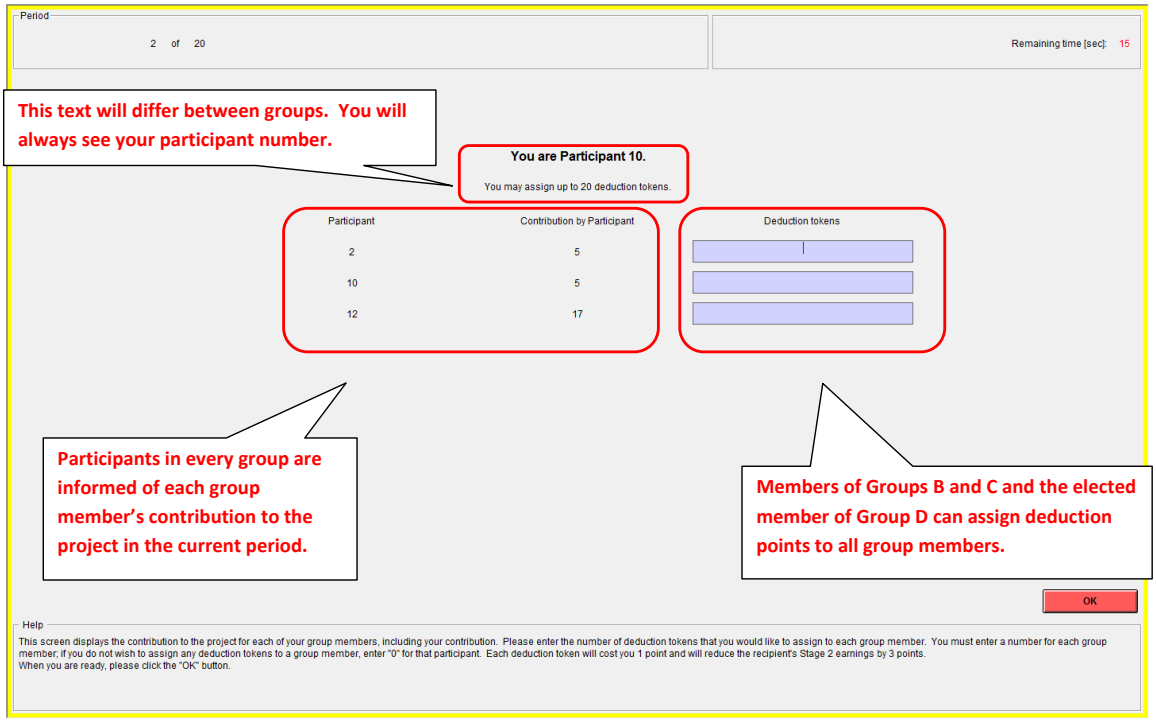
Stage 2: Contribution

Stage 2 is the same in all groups; participants receive 20 points and make contribution decisions on the basis of the same rules as in Part I of the study. In Group C and Group D, group members will also be informed of the group opinion.



Stage 3: Deduction points

In Stage 3, all participants receive 20 points and are informed of every group member's contribution to the project in the current period. In addition, members of Groups B and C and the elected member of Group D can assign deduction points to all group members.



Please complete the questions on the next sheet. If you have any questions or are finished answering the questions, please raise your hand. When everyone has completed the questions correctly, we will begin Part II of the study.

The following table may be helpful when completing the questions.

1.5*1 = 1.5	1.5*6 = 9	1.5*11 = 16.5	1.5*16 = 24
1.5*2 = 3	1.5*7 = 10.5	1.5*12 = 18	1.5*17 = 25.5
1.5*3 = 4.5	1.5*8 = 12	1.5*13 = 19.5	1.5*18 = 27
1.5*4= 6	1.5*9 = 13.5	1.5*14 = 21	1.5*19 = 28.5
1.5*5 = 7.5	1.5*10 = 15	1.5*15 = 22.5	1.5*20 = 30

A.7.5 Experimental instructions: Part 3

Your task

In this task you have been randomly paired with another person, whom we will refer to as the other. This other person is someone you do not know and will remain mutually anonymous. All of your choices are completely confidential. You will be making a series of decisions about allocating points between yourself and another person. We will refer to this person as the other. For each of the following questions, please indicate the distribution you prefer most by clicking on the button below your preferred distribution. You can only make one choice for each question. Your decisions will yield money for both yourself and the other person. There are no right or wrong answers, this is all about personal preferences. As you can see, your choices will influence both the amount of money you receive as well as the amount of money the other receives. Your choices will not be revealed to any of the other participants. You will also not be informed of which choice determines your payment.

Your payment

You will be randomly paired with another participant in the study. One of your choices will be selected at random and paid out in cash, so that both you and the other person will receive a payment based on your decision. In addition, you will also be the recipient of the choice made by another randomly selected participant. This person will be different from the person who receives payment based on your decision. Each point is worth 0.10 CHF (eg. 10 points = 1 CHF). We will combine both of your payments into one lump sum and mail that amount to the address you provide at the end of the study.

Appendix B

What you see is what you get? The effect of facial cues on trust-related behavior

This chapter is being prepared for submission to a leading economics journal and follows standard formatting for such journals. Work in this chapter was conducted with Bastiaan Oud, Jan Engelmann, Eva Krumhuber, and Ernst Fehr. This chapter was written by Tony Williams.

B.1 Abstract

Economic and social life are dominated by encounters with strangers whom we may never come into contact with again, and we need to determine which of these strangers to interact with and which ones to avoid. Given these encounters, people need to somehow discriminate among unknown individuals and select partners based on "first impressions." While it seems almost preposterous that we would choose to interact with people we believe are untrustworthy, relatively little is know about how we determine which individuals are trustworthy and therefore deserving of our trust. This paper experimentally examines the basis of these beliefs using static facial features by allowing participants in a trust game to see photographs of their counterparts, relying on both subjective ratings and an exogenous manipulation. By weakly relaxing anonymity in a lab experiment, this paper is able to identify a distinction between dispositional reciprocity and type-dependent reciprocity. Moreover, a large number of people react strongly to the perceived information in photographs and condition their behavior on these perceptions yet, paradoxically, these perceptions are wholly uninformative.

B.2 Introduction

The need to judge trustworthiness of strangers will little or no information is pervasive in social and economic interactions. Within firms, monitoring is costly and workers often have opportunities to shirk, requiring firms to trust employees, so firms may prefer to hire trustworthy workers. Firms may possess private information about profits, and the perceived trustworthiness of managers is likely a key determinant in the firm's ability to convince workers of the need to renegotiate contracts. Consumers may be willing to pay higher prices to a well-known online retailer to offset potential feelings of betrayal aversion if a package does not arrive (Bohnet and Zeckhauser, 2004). More generally, we make many decisions each day about who we interact with, and these decisions often need to be based on "first impressions" in the absence of additional information. These

interactions are not always repeated, which can prevent concerns for building a good reputation from offsetting selfish tendencies. In situations where interaction is repeated, allowing for the formation of reputation, the initial perceptions still serve as prior beliefs which are updated based on later interactions, and more positive information is needed to establish a good reputation when the prior suggests that the partner is not trusting or trustworthy. Despite its importance, little is known about the basis for perceptions about others' tendencies to engage in trusting and trustworthy behavior and whether these perceptions are either accurate or relevant in incentivized settings.

This paper experimentally examines the effect of observable facial features on trust and trustworthiness. We restrict our attention to static photographs of subjects' faces in a modified version of the trust game (Berg et al., 1995). While a greater understanding of static and dynamic features and their roles in determining perceptions of trustworthiness is desirable, we feel that this restriction is important experimentally and for many real-world settings outside the lab. From a practical perspective, we are able to maintain the greatest experimental control by using static photographs in which non-facial features are cropped out, thereby reducing or eliminating potential confounds. Perhaps more importantly, the widespread and increasing use of static photographs in online profiles related to employment, dating, and social networks (e.g. LinkedIn, Match.com, and Facebook) justifies our restriction and highlights its importance in naturalistic settings where social and economic decisions are made. Prior to online media, print advertisements for professional services, such as real-estate agents and lawyers, commonly did (and still do) use static photographs of the service provider.

In a broader context, our work is motivated by the evolution of altruism and cooperation, which has mostly been the domain of biology and psychology. Specific phenotypic features that signal willingness to cooperate are referred to as "green beards" in the evolutionary literature, though their existence and feasibility as an explanation for altruism and cooperation are often considered unrealistic by biologists; however, a small number of "green beards" have been found in other species (see thorough review in Henrich

(2004)). In spite of this view, a vast literature exists in psychology examining the ability of humans to predict the behavior of others based on static and dynamic features, such as bizogymatic width-to-height ratio in males (Stirrat and Perrett, 2010), difficult-to-fake features while smiling (Krumhuber et al., 2007), and "thin-slice" videos without sound (Vogt et al., 2013). However, even in cases where subjects can predict behavior better than chance, there is still little reason to believe that such skills are accurate enough to be plausible explanations for the evolution of altruism (Fehr and Fischbacher, 2005). We discuss our results with respect to "green beards" and as explanation for cooperative behavior in Section B.8; lack of repeated interaction rules out direct reciprocity as a motivation for behavior in our study (Trivers, 1971). Such identifiable features have been explored in economics, notably Frank (1987) (see also follow-up on evolutionary stability in Harrington (1989) and Frank (1989)).

In interpreting our results, we often refer to reciprocity, and it is important to clarify what we mean here. We rely on a definition common in economics, but not biology, that reciprocity is "being nice to those who are nice to us" and similarly for those who are not-so-nice. From the evolutionary and biological perspective, Trivers (1971) explains clearly that "[m]odels that attempt to explain altruistic behavior in terms of natural selection are models designed to take the altruism out of altruism." For instance, his own model of reciprocal altruism can be characterized precisely in terms of cartel models in industrial organization relying on standard folk theorems for repeated games (Tirole, 1988). Indirect reciprocity does not rely on repeated interactions with the same individual, but it does require more than one round of interaction and some knowledge about a counterpart's reputation; a model of indirect reciprocity that allows for uncertainty about the reputation of others can be relevant for trying to understand our findings, such as Panchanathan and Boyd (2003). For general overviews of evolutionary approaches to reciprocity, see Nowak and Sigmund (1998) and Nowak (2006) for a biological persepective and Sethi and Somanathan (2003) for an economic perspective.

Closest to our paper in the economics literature is Eckel and Petrie (2011), who also

use photographs of one's counterparts in a trust game. Our study differs from their paper in a few important ways. The primary difference between designs is that Eckel and Petrie (2011) showed photographs of counterparts who were currently in the lab. While there is some chance that subjects interacted while waiting prior to their experiment, this concern is relatively minor. A more pressing concern in face perception research is that non-facial features may be present in photographs (eg. clothing, jewelry, hair style and color), and these non-facial features may be the features utilized by subjects. While this concern does not affect interpretation of their results, the concern is relevant in our motivation. We avoid this concern by having our subjects interact with photographs of previous participants in an identical trust game (with the exception of photographs) but, prior to the experiment, we place an oval cut-out over the images to exclude non-facial information. We are also able to collect data on perceptions of trustworthiness and correlate these perceptions with behavior. Additionally, we have a treatment with an exogenous manipulation of trustworthiness. Overall, we show that these perceptions can be reliably predicted but are also inaccurate and uninformative (though, it should be noted, Eckel and Petrie (2011) also find no evidence that first-movers in a trust game can accurately infer second-mover behavior). Finally, we avoid the possibility of self-selected face-to-face interaction, which they identify as important to increase first-mover earnings, because only one party is present in the lab at any given time, and there is no need to pay money to view photographs in our study (similar to their control treatment in which there was no option and subjects always viewed a photograph).

While our primary motivation was to examine determinants of cooperation in the context of trust and trustworthiness, one of our key findings warrants an early discussion. Individuals form perceptions of trustworthiness from static photographs and use them in incentivized settings, yet these perceptions are, on average, incorrect and lead to earnings no better than random choice in our task. Goldin and Rouse (2000) examine the effect of blind auditions on hiring by orchestras and finds that women are more likely to be hired under blind auditions than non-blind. The uninformative nature of perceptions

from static photographs also calls into question the common practice in many European countries of including a personal photograph on resumés, which causes similar concerns to the role of discrimination towards stereotypically "black" names on resumés (Bertrand and Mullainathan, 2004). Both of these other papers highlight the (mis)use of uninformative signals as being informative, leading to discrimination in labor markets. Our paper cannot address this broader theme, but it is relevant for research on discrimination in general and implicit discrimination more specifically (Bertrand et al., 2005; Chugh, 2004; Greenwald et al., 1998).

Our findings also suggest an important dimension of discrimination in dynamic settings, resulting in a self-fulfilling prophecy that strengthens stereotypes. In the example we present here, the initial perception is based on race, but it may be any other observable feature (and may, for instance, be a perception regarding a patron's dispositional generosity). For instance, a widespread stereotype in the United States is that African-Americans are poor tippers in restaurants. In anticipation of a small tip, servers may divert effort to other patrons from whom they expect higher tips, thereby providing poor service to African-Americans. If tips reflect the quality of service received, one would then expect lower tips from African-Americans due not to their race but to poor service received in anticipation of a low tip (Conlin et al., 2003). Evidence suggests that such discrimination does occur and likely leads to self-reinforcing stereotypes (Brewster and Rusche, 2012; Dirks and Rice, 2004; Noll and Arnold, 2004).

Finally, our paper helps fill a gap in the literature between general altruism, such as pure or warm-glow altruism (Andreoni, 1989, 1990, 1995), and targeted altruism with the hope of future gains from interaction (Leider et al., 2009). The latter form of altruism is direct reciprocity in the evolutionary sense; see footnote 1 and Trivers (1971).

The remainder of the paper proceeds as follows. The modified trust game used in this study is described in Section B.3. Our experimental treatments are described in Section B.4. Empirical specification and relation to existing models is in B.5, followed by results and robustness in Sections B.6 and B.7, respectively. Discussion of the results is

in Section B.8, and we end with concluding remarks.

B.3 Modified Trust Game

We use a modified version of the trust game developed by Berg et al. (1995). The game consists of two players, Player A and Player B. Both players receive the same initial endowment, e, denoted in Swiss Francs (CHF). Player A has a binary decision and can send $x \in \{0, m\}$ to Player B. The amount sent by Player A is then tripled by the experimenter, and this tripled amount is given to Player B. Player B can then send any amount $y \in \{0, 1, \ldots, 3x\}$ back to Player A. Final monetary payoffs are given by

$$\pi_A = e - x + y \tag{B.1}$$

$$\pi_B = e + 3x - y \tag{B.2}$$

We elicit Player B decisions using the strategy method. As a result, Player B only has to give a conditional back transfer for the case when $x = m$. When $x = 0$, Player B cannot send any back transfer. This information was provided clearly both in the instructions and on the decision screen for Player B. We emphasize this detail here and elsewhere in the paper because it is important when interpreting our results; importantly, it rules out confusion about the conditional nature of the decision as an explanation for Player B decisions (i.e. it is implausible that Players B thought that they were making an unconditional back transfer). The decision screen for Player B contained a table in the following form (translated from the original German text; see Appendix B.11 for actual screen shots):

If Player A has sent the following amount over:	You will receive:	How much would you like to send back to Player A?
0	0	0
10	30	(empty space for entry)

In our experiment, we set $e = 12$ and $m = 10$. Both players received the same endowment to minimize concerns about inequity-aversion as a motivation for Player A to send positive amounts of money to Player B due to inequality of endowments. Moreover, we set $m = 10$ instead of $m = 12$ to prevent an all-or-nothing decision by Player A that could lead either or both players to extreme decisions (eg. Players A unwilling to risk everything by trusting or Players B feeling a stronger need to send back a non-zero amount). In the remainder of the paper, we refer to subjects in the role of Player A as first-movers and subjects in the role of Player B as second-movers and their decisions as transfers and back transfers, respectively.

B.4 Experimental Treatments

In all experimental sessions, subjects were recruited online using ORSEE (Greiner, 2004) and sessions were conducted in a computer lab at the University of Zurich. All experiments are programmed using z-Tree (Fischbacher, 2007). Participants were primarily students from the University of Zurich and Swiss Federal Institute of Technology (ETH Zürich). Individuals studying economics or psychology were excluded from recruitment. Over all treatments, two subjects are excluded from analysis. One subject had two accounts in ORSEE and participated in both an initial session and a treatment UNMODIFIED session (after which he informed the experimenters that he viewed his own photograph). One subject in treatment MODIFIED was asked to leave for using a laptop and delaying the experimental session. Details on number of subjects and average earnings for all treatments are presented in Table B.1. In all sessions, subjects received a show-up fee of 10 CHF for participating in the study, in addition to any earnings obtained during the experiment.

Treatment	N	Trustors	Trustees	Average earnings (CHF)
Initial session	84	42	42	28.67
UNMODIFIED	116	55	60	29.13
MODIFIED	157	75	81	29.67
Panel (oval cutouts)	174	—	—	33.41
Panel (no oval cutouts)	92	—	—	33.34
Guess modifications	29	—	—	43.59

Table B.1: Summary statistics

B.4.1 Initial Sessions

In research focusing on facial features, the standard practice is to crop the face using an oval cut-out to remove non-facial features, such as hairstyle and earrings, since these features may influence perceptions about the person in the photograph in ways unrelated to the research question. We chose to follow this standard practice, which required us to collect photographs in advance of the main treatments. Additionally, we employ face morphing technology in one of our treatments, which also required collecting photographs in advance. This approach allows us to exclude non-facial features as an explanation for our results; nonetheless, these other features may be relevant in forming perceptions and making decisions, and exploration of these other features remains an open and interesting question for future research.

Subjects were invited to the lab and played five rounds of the modified trust game described above ($N = 84$, average earnings = 28.67 CHF). Subjects were randomly matched with a new partner every period. In these sessions, subjects did not see photographs of their counterparts. Prior to the start of the experiment, subjects were informed that there would be a second optional study that they could participate in and that they would receive details on this second study later.

After completing the trust game, an experimenter publicly announced the optional study while consent forms were distributed to all subjects. Subjects were told that they would receive 10 CHF if they allowed us to take a photograph of their face and use it in subsequent studies. They were also told that they may potentially earn additional money

from the use of their photograph and strategies in future studies and that any earnings would be mailed to them.

All subjects then began completing a series of long questionnaires. We then proceeded in a specific manner to minimize selection bias. During the questionnaires, each subject was individually taken to a side room where a camera was set up for taking photographs. At this time, subjects gave us their consent form indicating whether they wanted to participate or not. After submitting the consent form (and, if applicable, having their photograph taken), subjects returned to the main lab and continued completing the questionnaires. As a result, each subject informed the experimenters in private whether or not they chose to participate, and these decisions could not influence other participants. Each subject was outside the main lab for approximately two minutes, making it difficult for subjects to infer the participation decisions of others. We also used the long questionnaires so that all subjects would be kept busy with the main experiment until after each subject was given the opportunity to participate in the second study, as otherwise some subjects may have finished early and chosen to leave the lab before having the option to participate in the second study. After completion of the questionnaires, one round was randomly selected for payment; the experimental currency was denoted in Swiss Francs, so there was no need for an exchange rate.

Sixty-three subjects (75%) allowed us to take their photograph and use their photograph and behavior in the lab in future studies. We find no significant differences in behavior between those who allowed us to take their photographs and those who did not (Kolmogorov-Smirnov test for equality of distributions: first-movers, exact $p = 0.208$; second movers, exact $p = 0.735$). We excluded one photograph in our main treatments in order to balance the number of first-movers and second-movers and to balance gender (31 photographs for each role; 17 male first-movers, 15 male second-movers).

B.4.2 Main Treatments

In our main experiments, new subjects were provided instructions for the modified trust game and informed that (i) they would see a photograph of their counterpart and (ii) these counterparts previously came to the lab and played the same game without seeing photographs of the other players. Subjects were also informed that their decisions would affect both their own payoffs and the payoffs of their counterparts, as the counterparts would receive their payment by mail. At the start of the experiment, each subject was assigned to the role of either Player A or Player B and remained in this role for the duration of the experiment. Each counterpart that provided a photograph in our initial sessions is seen once by each subject in the opposite role in our main experiments, for a total of thirty-one periods.

In treatment UNMODIFIED ($N = 116$, average earnings = 29.14 CHF), subjects saw the original photograph (after cropping) of their counterpart in each round. The middle photograph in Figure B.1 is an example. In treatment MODIFIED ($N = 157$, average earnings = 29.67 CHF), each photograph was morphed with computer-generated prototype faces that have previously been shown to be perceived as "more trustworthy" or "less trustworthy" in non-incentivized settings (Oosterhof and Todorov, 2008). Examples of the prototype faces can be seen in the far left and far right in Figure B.1. The middle left and middle right photographs in Figure B.1 show the versions of the original photograph used in our MODIFIED treatment. Each subject in our MODIFIED treatment saw each counterpart once, and half (either 15 or 16) of the images were morphed to be more trustworthy and the remaining images morphed to be less trustworthy. Morphing was done using JPsychoMorph software (Tiddeman et al., 2005).

Photographs were presented in random order in our UNMODIFIED treatment. In our MODIFIED treatment, the modified images were counterbalanced across subjects, and we imposed a requirement that one type of modification could not be seen more than six times in a row. We did this to reduce the chance of subjects noticing commonalities among the morphed images. A concern still remains that subjects may be able to detect

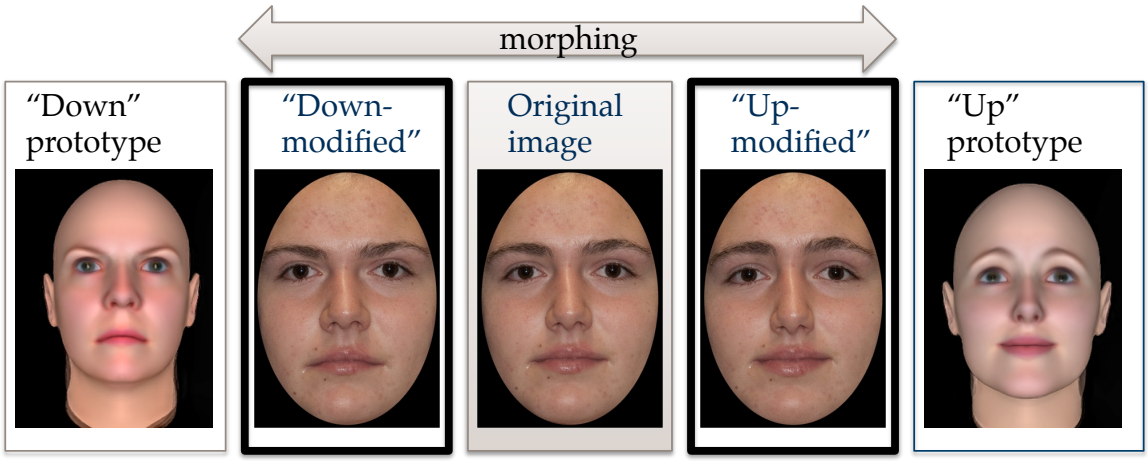


Figure B.1: Example of photographs and computer-generated faces. Middle represent and unmodified image. Far left and far right are examples of computer-generated faces that are claimed to be less- and more-trustworthy, respectively (Oosterhof and Todorov, 2008). Each computer-generated face was morphed with the unmodified photograph to create one downmodified image (middle left) and upmodified image (middle right).

that images have been modified. [1]

After completing thirty-one rounds of the modified trust game, subjects then saw the images in the same order and were asked to rate them based on a series of characteristics, including trustworthiness. Photographs were presented in the same order to maximize the amount of time between subjects' decisions in the trust game and their ratings and to reduce the chances of a consistency bias between behavior in the experiment and subjective ratings. Additional details on these ratings will be provided below in our results in Section B.6. These ratings provide us with a subjective measure of perceived trustworthiness in both our UNMODIFIED and MODIFIED treatments, while our exogenous

[1] An independent sample of subjects ($N = 29$, average earnings = 43.59 CHF) was recruited to examine whether or not the morphed images were detectable. In the first part of the experiment, these subjects saw two photographs from each of the 62 subjects in our initial session who provided consent (114 photographs); one photograph was image used in our UNMODIFIED treatment, and the other was one of the two images used in our MODIFIED treatment. These subjects also saw an additional 16 photographs morphed in a similar manner but which did not contain oval cutouts in a second part. In the experiment instructions, subjects were informed that 10 of the 140 total rounds from both parts would be randomly selected; each accurate guess would add 4 CHF to the subject's earnings, and each incorrect guess would add 0 CHF to the earnings. They were able to guess whether an image had been morphed only slightly above chance (56.8% over the first 114 periods) and did not show any learning effects at the individual or aggregate level; the marginal effect of $Period$ in a probit regression, $Correct = \beta_0 + \beta_1 Period + \varepsilon$, again only using the first 114 periods, is 0.0002 ($p = 0.25$, robust standard errors clustered by subject). As a result, we are reasonably confident that our morphed photographs in treatment MODIFIED do not appear unnatural and are not easily distinguished from the unmorphed images used in the same sessions.

manipulation in the MODIFIED treatment gives us an objective measure (validated by an external sample, as described later in Section B.7).

Following the ratings, subjects completed a short questionnaire; afterwards, one round was randomly selected for payment; the experimental currency was denoted in Swiss Francs, so there was no need for an exchange rate.

B.5 Empirical Specification

Before proceeding to our results, we need to specify a functional form to gain tractability and formulate predictions. With first-movers, there is a strategic motive to assess trustworthiness of one's counterpart; however, we cannot disentangle preferences over final outcomes from beliefs about expected back transfers from second-movers. Additionally, first-movers make a binary decision, leaving a probit (or logit) as the only feasible estimation option, and we can include subjective trustworthiness ratings. For first-movers, we are interested only in whether they discriminate in their trusting decisions based on their perceptions about their counterparts' trustworthiness.

We instead focus more on model specification for second-movers, as second-movers do not face any strategic motive to send money to first-movers and act essentially as dictators in a dictator game; nonetheless, substantial evidence exists demonstrating that second-movers will often send a back transfer. Our final estimation will use the same linear specification as for first-movers, but it has an intuitive model-based interpretation for second-movers.

Importantly, we can decompose second-mover behavior into dispositional reciprocity and type-based reciprocity. Dispositional reciprocity corresponds to traditional models of inequity aversion and is a preference over final states of wealth independent of the characteristics of one's counterparts (aside from wealth). Type-based reciprocity corresponds to differential back transfers based on perceptions of a counterpart's average behavior and is captured intuitively in Levine (1998). Because second-movers only send conditional

back transfers (since we use the strategy method), any discrimination in back transfer behavior cannot be based on expectations about receiving money from the first-mover.

For a variety of reasons discussed below, we opt for a simple Cobb-Douglass specification for second-movers,

$$U(\pi_S, \pi_O) = (\pi_S)^{1-\alpha-\beta T}(\pi_O)^{\alpha+\beta T} \tag{B.3}$$

where π_S and π_O denote payoffs to self and other, respectively, and T indicates the perceived trustworthiness of the other player. We require $\alpha, \beta \in [0, 1]$ and, for technical reasons, also require $T \leq (1 - \alpha)/\beta$. If we let $\pi_S + \pi_O = w$ and use the substitution $\pi_S = w - \pi_O$ in equation (B.3), maximizing the utility function yields the standard Cobb-Douglass condition:

$$\pi_O = (\alpha + \beta T)w, \tag{B.4}$$

so that the share of total wealth being given to the first-mover is determined by the second-mover's dispositional reciprocity (α) and type-based reciprocity (β). Note that, since the second-mover only makes a conditional back transfer when the first-mover trusts, the actual decision of the first-mover does not factor into the second-mover's back transfer decision; in addition, we can also see that $w = 2(e + m)$.

The first major reason for specifying Cobb-Douglass preferences is its consistency with Andreoni and Miller (2002)'s findings on altruistic preferences. They specify a constant elasticity of substitution (CES) model, of which Cobb-Douglass is a special case. In their study, they use a modified dictator game and vary both endowments and relative prices of giving money to the other person, allowing them to estimate elasticity of substitution. We keep the relative price of giving fixed, preventing estimation of the elasticity of substitution. Cobb-Douglass allows for two of the three main preference classes identified by Andreoni and Miller (2002), selfish and Leontief (perfect equality), while our design does not provide specific predictions for their third class, perfect substitutes. Any Player A who has perfect substitute preferences or maximizes social efficiency will always send

$x = m$, as this increases the total payoff to both players from $2e$ to $2(e + m)$.

The second reason for using Cobb-Douglass utility relates to the array of social preferences models in the literature (Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Fehr and Schmidt, 1999). The most common models are linear and make extreme predictions in two-person games, predicting either perfect selfishness or perfect equality. Our Cobb-Douglass specification allows for both of these predictions as well as intermediate levels of inequality. Selfish behavior is consistent with $\alpha = \beta = 0$, while perfect equality is captured by $\alpha = 0.5$ and $\beta = 0$.

Third, we want to allow for players to discriminate in their decisions based on subjective perceptions of their counterparts. Such discrimination is captured by the logic of Levine (1998). In his model, subjects respond not to the actual behavior of others but to a perception about how the other person generally behaves. The model would predict perfect selfishness in our experiment, but his notion of type-based altruism is captured in our specification by $\beta > 0$.

B.6 Results

B.6.1 Behavior in main treatments

We begin by analyzing first-mover behavior, and our first major result is not surprising given the strategic motive for first-movers to make inferences about the trustworthiness of second-movers.

Result B.1. *First-movers discriminate based on perceived trustworthiness in both UNMODIFIED and MODIFIED treatments. Effect sizes for perceived trustworthiness are not significantly different between treatments and are nearly identical.*

The result can be seen in the coefficients for subjective ratings (Trustworthiness) in the top row of Table B.2 across the MODIFIED, UNMODIFIED, and pooled data. In all cases, an one-unit increase in the perceived trustworthiness leads to about a 10% increase

	Unmodified	Modified			Pooled
	(1)	(2)	(3)	(4)	(5)
Trustworthiness	0.097	0.101		0.099	0.099
	(0.026)***	(0.018)***		(0.019)***	(0.015)***
Upmodified			0.108	0.021	
			(0.021)***	-0.023	
N	1705	2325	2325	2325	4030

Table B.2: First-mover transfer decisions. Probit regressions (marginal effects reported). Robust standard errors clustered by subject. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

in the probability of transferring and is highly significant ($p < 0.01$).

Our manipulation in treatment MODIFIED also provides an exogenous source of variation and allows us to test the credibility of claims regarding trustworthiness perceptions from facial features.

Result B.2. *Perceived trustworthiness varied significantly from unmodified photographs in both predicted directions in treatment MODIFIED. Moreover, subjects were 11% more likely to trust the upward-modified photographs over the downward modified photographs.*

The increase of 11% in trusting behavior can be seen in column (3). Interestingly, the effect shifts completely to subjective trustworthiness ratings, which can be seen in column (4); the effect of perceived trustworthiness is still about 10% and highly significant, but the effect of the modification is now only about 2% and is no longer significant. Thus, the modification appears to be working directly through changes in subjective perceptions of trustworthiness.

In light of our first two results, it is apparent that subjects are indeed making inferences about the trustworthiness of their counterparts and discriminating based on these inferences. If these inferences are valid, subjects should have high earnings in our study. Our next result contradicts this hypothesis.

Result B.3. *Average earnings for first-movers in both treatments are not significantly different from choosing a strategy at random and are significantly lower than if subjects*

Strategy	Mean	Std. Error	[95% CI]	
Best response	13.95	0.08	13.80	14.10
Never trust	12.00	—	—	—
Actual earnings	10.42	0.13	10.16	10.68
Random choice	10.42	0.09	10.25	10.59
Always trust	8.83	0.17	8.49	9.18
Number of observations: 1705				

Table B.3: Expected earnings from different strategies and actual choices for UNMODIFIED treatment. Our MODIFIED treatment is omitted from analysis to avoid potential bias introduced by morphing photographs. Best response assumes subject could maximize payoff on each trial. Subjects could expect to earn, based on their actual choices, no more than if they decided randomly.

ignored photographs and never trusted second-movers.

Table B.3 presents average earnings for first-movers and places them in the context of other potential strategies. Actual average earnings were 10.42 CHF. Subjects using a fixed strategy of never trusting would earn 12 CHF with certainty, while always trusting would have led to earnings of 8.83 CHF on average. Therefore, in our sample, trusting does not pay off on average, in line with previous trust games (Camerer, 2003). However, a subject who is able to accurately predict her counterpart's back transfer on each trial would earn 13.95 CHF on average, making the ability to detect the trustworthiness of others a profitable skill. Actual earnings are significantly less than the hypothetical maximum earnings and earnings from a fixed strategy of never trusting but are also significantly higher than a fixed strategy of always trusting. Importantly, we can also compute the expected payoff from random choice, in which a subject trusts with probability 0.5. Random choice would yield expected earnings of 10.42 CHF, which is identical to our subjects' actual average earnings.

Result B.4. *Second-movers also appear to discriminate based on perceived trustworthiness. However, second-movers can be divided into roughly four groups. First,* 40% *of subjects are purely selfishly and always send zero. Second,* 27% *of subjects display dispositional reciprocity but not type-dependent reciprocity. Third,* 10% *of subjects display*

type-based reciprocity but not dispositional reciprocity. Finally, 18% *of subjects display both dispositional and type-dependent reciprocity. Seven subjects (*5%*) cannot be categorized.*

We begin by estimating the Cobb-Douglass preferences introduced previously in Section B.5 and use the linear specification

$$(\text{Back Transfer}) = \alpha + \beta(\text{Trustworthiness}) + \varepsilon \tag{B.5}$$

where α measures dispositional reciprocity and β captures type-based reciprocity. Due to potential censoring, we use a two-sided Tobit regression. Figure B.2 displays individual-level parameter estimates, and we set non-significant parameters to zero. Roughly 40% of subjects are not significantly different from purely selfish preferences ($\alpha = 0, \beta = 0$). Another 27% of subjects display some degree of dispositional reciprocity but not type-dependent reciprocity ($\alpha > 0, \beta = 0$); these subjects' preferences conform to the logic of outcome-based models of inequity aversion. Next, 10% of subjects display only type-based reciprocity ($\alpha = 0, \beta > 0$); this behavior is consistent with the phenomenon captured in Levine (1998). Finally, 18% display both dispositional reciprocity and type-based reciprocity ($\alpha > 0, \beta > 0$), suggesting a more complex mix of preferences including both inequity aversion and type-based altruism. The distribution of subject classifications are listed in Table B.4.

A potential concern here is that second-movers were confused about the conditional back transfer and were engaging in reciprocity based on expected transfer. We again emphasize that the conditional nature of the back transfer was made explicit on the decision screen each time the subjects entered their decisions (see on-screen text in Section B.3 above and Appendix B.11 for experiment instructions containing screenshots).

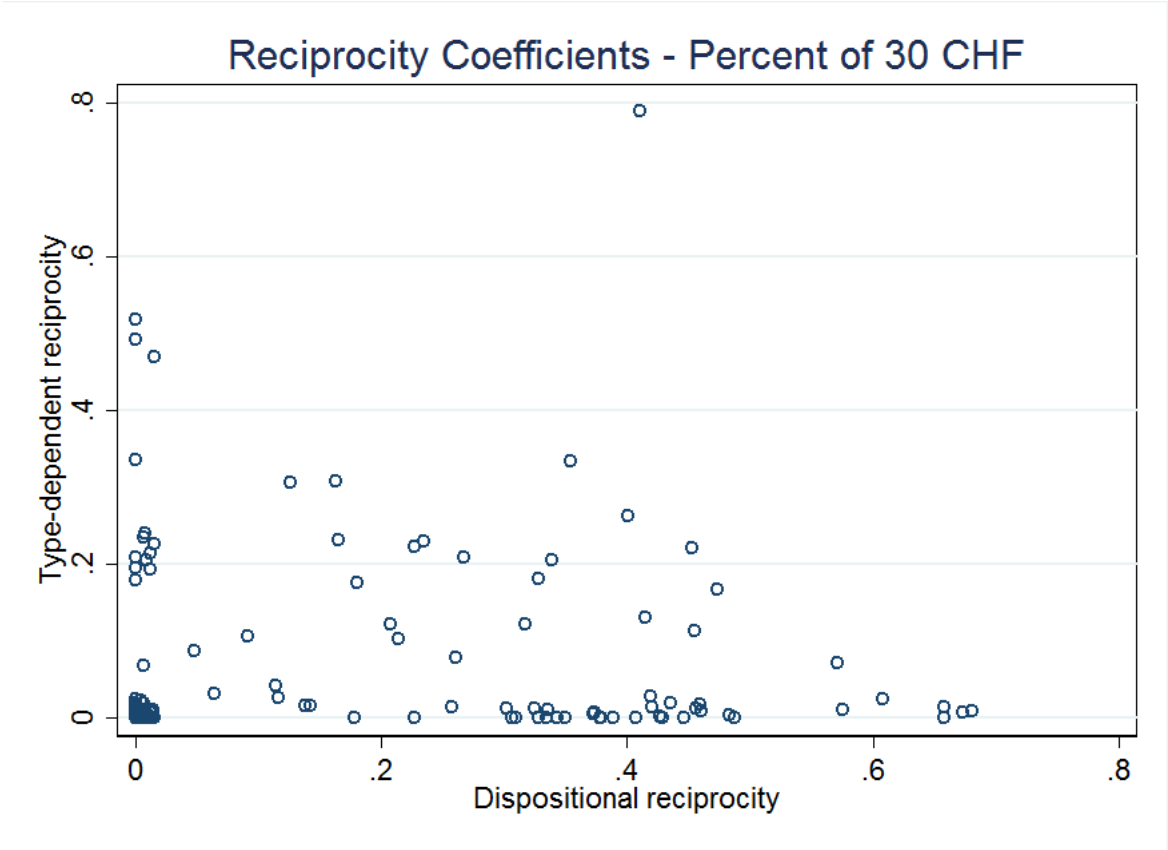


Figure B.2: Distribution of dispositional (α) and type-based (β) reciprocity parameters. Non-significant parameters are set to zero. Seven outliers are omitted; these correspond to the uncategorized subjects in Table B.4. Total payoff for both subjects is 44 CHF. Natural focal points for dispositional reciprocity include 0 (selfish), 0.33 (10 CHF, perfect repayment), 0.5 (15 CHF, equal split of amount received by second-mover), and 0.67 (20 CHF, equal final payoffs after including initial endowments). Slight jitter added so that all data points can be seen.

Type	Frequency	Percent
Selfish	57	40.43
Dispositional Only	38	26.95
Type-based Only	14	9.93
Dispositional and Type-based	25	17.73
Uncategorized	7	4.96

Table B.4: Subject classification into reciprocity types based on estimates from equation (B.4). Non-significant estimates are set to zero. Categories correspond to (α, β) parameters: Selfish ($\alpha = 0, \beta = 0$); Dispositional only ($\alpha > 0, \beta = 0$); Type-based only ($\alpha = 0, \beta > 0$); Dispositional and Type-based ($\alpha > 0, \beta > 0$); and Uncategorized ($\alpha < 0$ or $\beta < 0$ or both).

B.6.2 Predicting Behavior

Thus far, we have provided evidence that individuals form perceptions of trustworthiness based on certain facial characteristics and that these characteristics can be exogenously manipulated to change trusting and trustworthy behavior in reliable, predictable ways. However, in all of these situations, decisions made by our subjects affect both their own payoffs and the payoffs of others. Therefore, subjects' decisions reflect motives such as efficiency or inequity-aversion and do not accurately reveal beliefs about the general behavior of counterparts. To address these concerns, we conducted additional sessions with new independent samples. In these sessions, which we refer to as PANEL, subjects did not play the modified trust game. Instead, we explained the trust game in the instructions as before, and subjects were incentivized to predict the average behavior of the person in the photograph. These new subjects always viewed unmodified images. In some sessions ($N = 174$, average earnings = 33.41 CHF), subjects saw the same images as in treatment UNMODIFIED, in which the photographs contained an oval cut-out. Since these images may seem artificial in some sense, we conducted additional sessions ($N = 92$, average earnings = 34.34 CHF) in which subjects viewed images without the oval cut-out. The important aspect here is that decisions of subjects in our PANEL sessions do not affect the payoffs of others, which rules out concerns about efficiency, reciprocity, and inequity-aversion. When viewing Players A, subjects were asked to predict what percent

of the time (0-100%) the person in the photograph transferred 10 CHF to his or her counterpart. When viewing Players B, subjects were asked to predict the average back transfer (0-30 CHF) the person in the photograph sent to his or her counterpart. The experimental currency in PANEL sessions is points. Subject earnings, in points, were $200 - 2 \times |\text{predicted transfer} - \text{actual transfer}|$ if the trial selected for payment contained a Player A and $200 - 5 \times |\text{predicted transfer} - \text{actual transfer}|$ if the trial selected for payment contained a Player B. The differing multiplicative terms were selected in advance to equalize expected earnings regardless of whether the trial contained a Player A or a Player B. At the end of the experiment, one round was randomly selected for payment; the experimental currency was denoted in points, with an exchange rate of 10 points = 1 CHF.

Subjects viewed two blocks during the experiment, with each block consisting of only first-movers or only second-movers. Each block contained 9-11 photographs, counterbalanced across subjects so that we had approximately the same number of predictions for each photograph, for a total of 18-22 predictions per subject. All subjects predicted behavior of both first-movers and second-movers. We used these blocks to avoid confusion resulting from full randomization of photographs and roles across trials, though photographs were randomized with a block. Order of block (first- or second-movers) was also counterbalanced across subjects. After making their predictions, subjects then viewed the photographs in the same order again and provided ratings on the same characteristics from the MODIFIED and UNMODIFIED treatments. These sessions provide our next major result.

Result B.5. *Subjects cannot predict average behavior of individuals in static photographs better than chance. Moreover, confidence in predictions does not increase accuracy of predictions.*

Our results for predictions can be seen in Table B.5. We find no evidence of accuracy at the aggregate level. It remains possible that a subset of the population can accurately infer trustworthiness from facial features or that it is easier for everyone to infer trust-

worthiness from some faces but not others. We also hypothesized that, if a subset of the population does in fact possess this skill, these individuals would be more confident in their predictions; therefore, we elicited confidence in prediction on each trial. Figures B.3 and B.4 demonstrate our (null) results intuitively. We find no evidence that a subset of the population can accurately infer trustworthiness from facial features or that it is easier for everyone to infer trustworthiness from some faces but not others when analyzing the data at the level of individual PANEL subject or individual photograph. Also, we find no evidence for accuracy of predictions increasing with confidence; people who are confident in their predictions are no more accurate than people who lack confidence in their predictions.

B.7 Robustness and Additional Measures

Our analysis of first-mover and second-mover behavior has been fairly simple. We have only used self-reported trustworthiness ratings as a proxy for trustworthiness perceptions, though our exogenous manipulation also provides evidence that there is a coherent basis for these perceptions. For now, we will restrict attention to our MODIFIED treatment, since it is an exogenous source of variation in trustworthiness that we may be able to explain away using additional measures.

Our definition of trustworthiness has been behavioral-based, though a more common but ill-defined notion is that trustworthiness is a personality trait, and we may potentially develop an index of trustworthiness using alternative measures. Along with ratings of trustworthiness, we also collected subjective ratings on the following features: attractiveness, anger, fear, happiness, surprise, disgust, sadness, dominance, masculinity, femininity, competitiveness, friendliness, and warmth. Given that several of these measures seem to be capturing aspects of the same latent characteristics and including all of these ratings as regressors in a single model would create a multicollinearity problem, we used factor analysis reduce these ratings into fewer distinct factors which can then be

used as regressors. Using the standard practice of retaining only factors with an eigenvalue greater than one, we are able to reduce our ratings to four factors, which we label as (i) positive affect, (ii) negative affect, (iii) power, and (iv) femininity. Details on factor loadings are in Table B.6.

	Actual transfer (% endowment)			Actual back transfer (CHF)		
	(1)	(2)	(3)	(4)	(5)	(6)
Prediction	0.030	0.098	0.051	-0.062	-0.036	-0.052
	(0.032)	(0.051)*	(0.028)*	(0.042)	(0.048)	(0.032)
Constant	65.719	62.482	64.766	6.865	7.21	6.974
	(1.634)***	(2.689)***	(1.413)***	(0.513)***	(0.553)***	(0.390)***
R^2	0.00	0.00	0.00	0.00	0.00	0.00
N	1798	951	2749	1798	950	2748

Table B.5: Panel accuracy. Raters in columns (1) and (4) saw photographs with oval cutout. Raters in columns (2) and (5) saw photographs without oval cutout. Columns (3) and (6) contain pooled data. OLS regressions. Robust standard errors clustered by subject. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

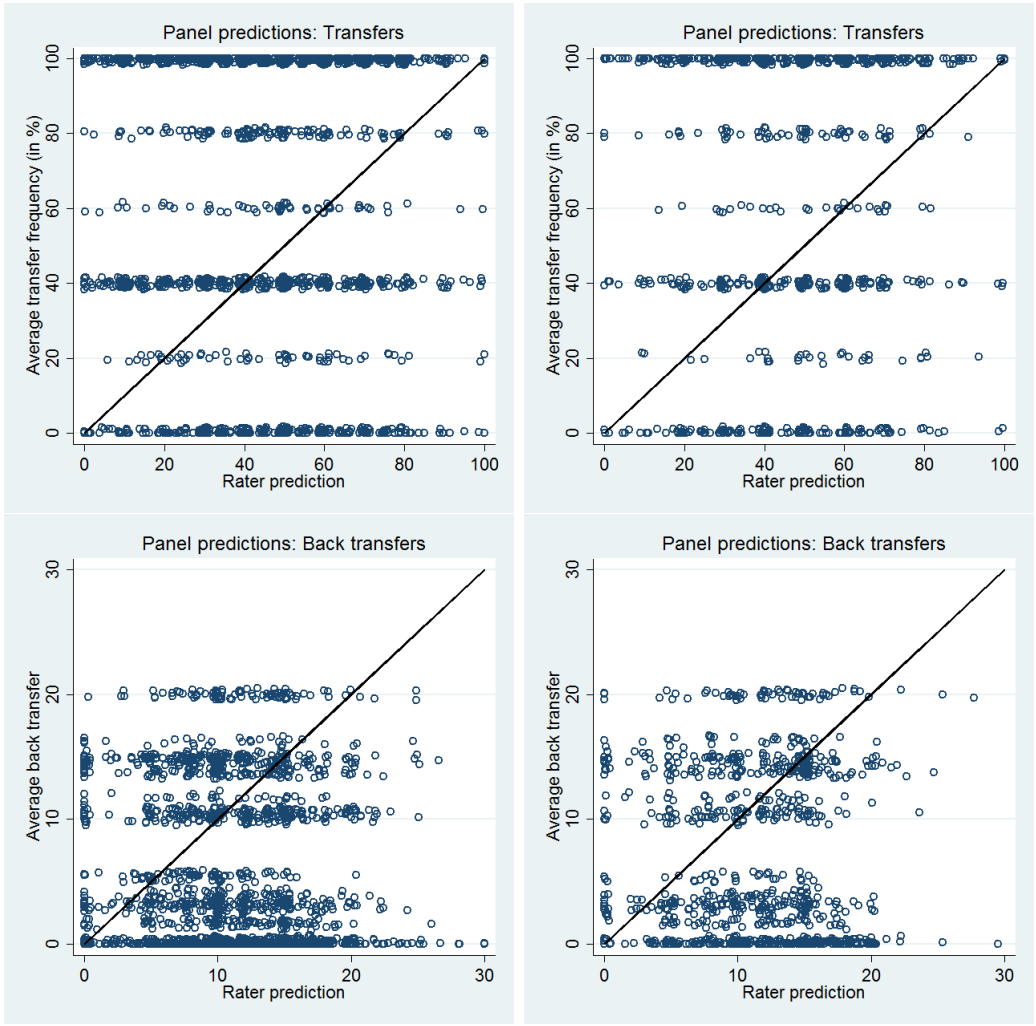


Figure B.3: Lack of accuracy in panel predictions. Images on left are from panel that did see oval cutouts; images on the right are from panel that did not see oval cutouts. Top row contains panel predictions when viewing first-movers; bottom row contains panel predictions when viewing second-movers. Slight jitter added so that all data points can be seen. The 45°-line is included for visualization purposes.

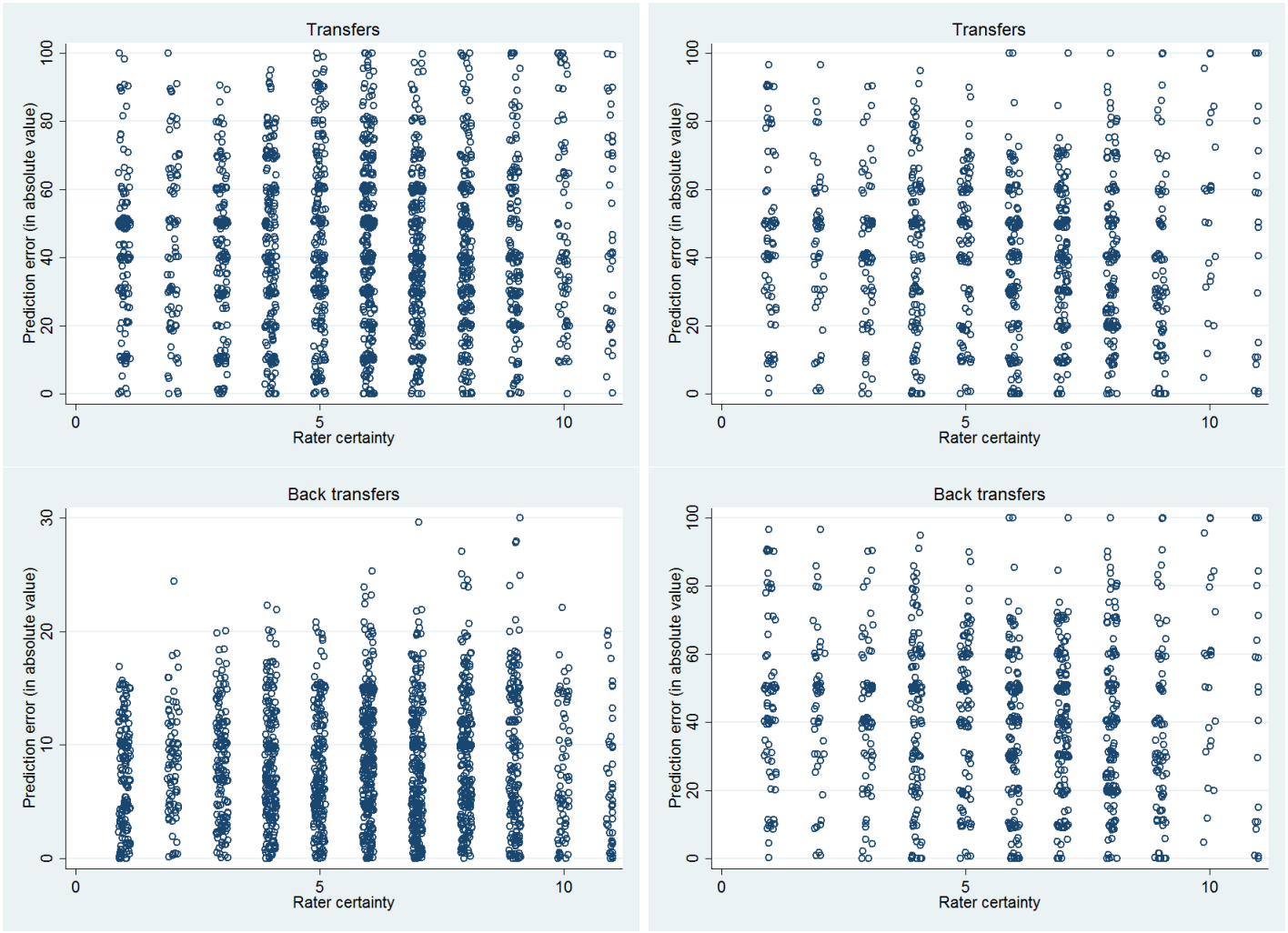


Figure B.4: Lack of accuracy in panel predictions as a function of rater certainty. Images on left are from panel that did see oval cutouts; images on the right are from panel that did not see oval cutouts. Top row contains panel predictions when viewing first-movers; bottom row contains panel predictions when viewing second-movers. Slight jitter added so that all data points can be seen.

Ratings	Factors				Uniqueness
	Positive affect	Negative affect	Power	Femininity	
Attractiveness	**0.60**	-0.02	0.36	0.26	0.44
Anger	-0.39	0.33	**0.64**	-0.12	0.31
Fear	0.13	**0.79**	-0.06	0.08	0.35
Happiniess	**0.83**	-0.05	0.03	0.07	0.30
Surprise	0.15	**0.68**	0.04	0.06	0.52
Disgust	-0.18	**0.64**	0.35	-0.04	0.43
Sad	-0.04	**0.73**	0.01	-0.03	0.46
Dominant	-0.11	-0.01	**0.87**	-0.07	0.23
Masculine	0.04	0.05	0.19	**-0.92**	0.10
Feminine	0.25	0.10	0.08	**0.89**	0.12
Competitive	0.19	-0.02	**0.78**	-0.05	0.36
Friendly	**0.89**	0.04	-0.12	0.08	0.19
Warm	**0.86**	0.08	-0.10	0.08	0.24

Table B.6: Ratings variables and factor analysis. After completing the trust game, subjects in treatment MODIFIED viewed the photographs in the same order and gave subjective ratings on the characteristics in the left-hand column. Ordering of variables was kept fixed for each subject but randomized across subjects. Factor loading greater than 0.50 in magnitude are marked in bold to indicate strong loadings. Factor analysis with principal components factors and varimax rotation.

We then drop trustworthiness ratings and include these factors as regressors to predict behavior for both first-movers and second-movers. We also include specifications with gender effects, as gender (of subject, counterpart, or both) seems an obvious a priori explanation for at least a portion of our trustworthiness perceptions. Results for first-movers are in Table B.7 and for second-movers in Table B.8.

Result B.6. *The effect of our exogenous manipulation in treatment MODIFIED on first-mover behavior can be explained by positive affect and femininity, while the back transfer decision of second movers in treatment MODIFIED is explained by positive affect.*

Surprisingly, we do not find robust evidence for any gender effect for either first-movers or second-movers. Importantly, the effect of our exogenous manipulation on perceived trustworthiness is no longer significant in any of our specifications after including our four factors capturing latent variables. In particular, positive affect and femininity are robust and highly significant in predicting transfers from first movers, while only positive

affect is robust and highly significant in predicting back transfers from second movers.

Even though we are able to decompose our other subjective ratings into these four factors for our sample with modified images, one may be concerned that (i) these factors do not capture our common notion of trustworthiness or (ii) trustworthiness perceptions are formed based on unnatural characteristics which are an artifact of our morphing procedure or based on aspects other than these four factors.

	(1)	(2)	(3)	(4)	(5)
Upmodified	0.108	-0.016	-0.011	-0.011	0.001
	(0.021)***	(0.025)	(0.026)	(0.025)	(0.025)
Positive affect		0.124	0.123	0.123	0.121
		(0.030)***	(0.029)***	(0.029)***	(0.029)***
Negative affect		-0.04	-0.038	-0.038	-0.041
		(0.036)	(0.037)	(0.037)	(0.037)
Power		-0.022	-0.02	-0.021	-0.019
		(0.029)	(0.03)	(0.03)	(0.03)
Femininity		0.052	0.05	0.061	
		(0.011)***	(0.025)**	(0.019)***	
Female Counterpart			0.028		0.098
			(0.047)		(0.027)***
Female			0.163	0.131	0.165
			(0.088)*	(0.088)	(0.089)*
Female*Female Counterpart			-0.068		-0.058
			(0.041)*		(0.041)
Female*Femininity				-0.027	
				(0.023)	
N	2325	2325	2325	2325	2325

Table B.7: Robustness of first-mover transfer decisions from MODIFIED treatment when including results from factor analysis and gender effects. We find no evidence of a robust gender effect. The effect of the upmodification disappears when including the four factors, suggesting that the modification primarily induces transfer by increasing perceptions of positive affect and femininity. Probit regressions (marginal effects reported). Robust standard errors clustered by subject. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

	(1)	(2)	(3)	(4)	(5)	(6)
Upmodified	0.671	-0.227	-0.124	-0.163	-0.128	-0.048
	(0.180)***	(0.311)	(0.297)	(0.302)	(0.307)	(0.169)
Positive affect		1.132	1.016	1.037	1.017	0.832
		(0.334)***	(0.338)***	(0.342)***	(0.340)***	(0.136)***
Negative affect		-0.805	-0.727	-0.714	-0.726	-0.008
		(0.426)*	(0.443)	(0.440)	(0.442)	(0.086)
Power		-0.011	-0.034	-0.051	-0.034	-0.198
		(0.306)	(0.313)	(0.311)	(0.313)	(0.086)**
Femininity		0.183	-0.033	0.007		0.2
		(0.183)	(0.299)	(0.349)		(0.075)***
Female Counterpart			0.573		0.532	0.212
			(0.402)		(0.319)*	(0.304)
Female			1.075	1.122	1.084	-3.322
			(1.371)	(1.393)	(1.351)	(0.512)***
Female*Female Counterpart			0.065		0.048	-0.065
			(0.445)		(0.407)	(0.425)
Female*Femininity				0.294		
				-0.402		
Constant	5.188	5.738	4.847	5.127	4.869	2.746
	(0.658)***	(0.699)***	(1.027)***	(1.076)***	(1.053)***	(0.348)***
R^2	0.00	0.04	0.05	0.05	0.05	0.81
N	2511	2511	2511	2511	2511	2511

Table B.8: Robustness of second-mover transfer decisions from MODIFIED treatment when including results from factor analysis and gender effects. We find no evidence of a robust gender effect. The effect of the upmodification disappears when including the four factors, suggesting that the modification primarily induces higher back transfers by increasing perceptions of positive affect. OLS regressions. Robust standard errors clustered by subject. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

To address these concerns, we generated predictions of the trustworthiness ratings given by the PANEL using only the four measures from our factor analysis and subsequent regression coefficients from treatment MODIFIED. We proceeded in four steps.

First, we used the factor analysis results to create factors for the PANEL by combining the PANEL ratings with the factor loadings from treatment MODIFIED. Next, we regress the trustworthiness ratings from our MODIFIED treatment on the factors generated using the ratings and factor analysis from our MODIFED treatment and form in-sample predictions using the following linear specification:

$$\begin{aligned}(\text{Trustworthiness Rating})_{MODIFIED} = \beta_0 &+ \beta_1(\text{Positive Affect}) \\ &+ \beta_2(\text{Negative Affect}) + \beta_3(\text{Power}) \\ &+ \beta_4(\text{Femininity}) + \varepsilon \end{aligned} \tag{B.6}$$

where the in-sample predicted trustworthiness for the MODIFIED treatment is

$$\begin{aligned}(\text{Trustworthiness Prediction})_{MODIFIED} = \hat{\beta}_0 &+ \hat{\beta}_1(\text{Positive Affect}) \\ &+ \hat{\beta}_2(\text{Negative Affect}) + \hat{\beta}_3(\text{Power}) \\ &+ \hat{\beta}_4(\text{Femininity}). \end{aligned} \tag{B.7}$$

Next, we take the same least squares estimates from the regression in equation B.7 and combine them with the out-of-sample factors using the PANEL ratings to generate

out-of-sample predictions of trustworthiness ratings by PANEL subjects:

$$\begin{aligned}(\text{Trustworthiness Prediction})_{PANEL} = \hat{\beta}_0 &+ \hat{\beta}_1(\widehat{\text{Positive Affect}}) \\ &+ \hat{\beta}_2(\widehat{\text{Negative Affect}}) + \hat{\beta}_3(\widehat{\text{Power}}) \\ &+ \hat{\beta}_4(\widehat{\text{Femininity}}) \qquad \text{(B.8)}\end{aligned}$$

.

Finally, we regress the trustworthiness ratings on these predictions,

$$\begin{aligned}(\text{Trustworthiness Rating})_{TREATMENT} = \alpha_0 &\\ &+ \alpha_1(\text{Trustworthiness Prediction})_{TREATMENT} \\ &+ \varepsilon \qquad \text{(B.9)}\end{aligned}$$

where $TREATMENT \in \{MODIFIED, PANEL\}$, and we additionally split the PANEL data into those who saw the unmodified images with an oval cut-out and those who saw the unmodified images without an oval cut-out. While there are valid reasons for using the oval cut-outs, as discussed previously, there is a potential concern that subjects view such images as appearing unrealistic or unnatural. The MODIFIED treatment serves as a baseline for comparison with the out-of-sample predictions.

Result B.7. *Our factors significantly predict trustworthiness ratings in our MODIFIED treatment. Moreover, out-of-sample trustworthiness predictions based on our MODIFIED treatment significantly and accurately predict the trustworthiness ratings from PANEL subjects using unmodified photographs.*

Results are presented in Table B.9. While we cannot rule out the possibility of other

	(1)	(2)	(3)
Trustworthiness	0.545	0.528	0.53
	(0.016)***	(0.013)***	(0.019)***
Constant	1.221	1.325	1.446
	(0.055)***	(0.043)***	(0.068)***
R^2	0.55	0.54	0.59
N	4836	3596	1901

Table B.9: Predicting trustworthiness ratings using factor loadings and OLS estimates from MODIFIED treatment. Column (1) is in-sample prediction from treatment MODIFIED. Columns (2) and (3) predict out-of-sample trustworthiness rating in the panels with and without cutouts, respectively. OLS regressions. Robust standard errors clustered by subject. $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

factors determining behavior, the results in Table B.9, along with the effect of subjective trustworthiness ratings on behavior (Table B.2 and Figure B.2), suggest that our modifications are in fact changing facial characteristics used to form perceptions of trustworthiness in a reliable and natural manner. In addition, recall that an independent sample with financial incentives was only slightly better than chance at determining whether or not a photograph was modified and did not show any learning effects, further suggesting that there are not unnatural facial features that are easily identified in our modified images (see Section B.4).

B.8 Discussion

Our results have at least one important implication for the interpretation of reciprocity (and altruism) in laboratory experiments. Typical trust games are anonymous interactions without feedback or additional information, so any non-selfish behavior demonstrated is necessarily interpreted as dispositional reciprocity. Among our sample, 55% demonstrate some form of reciprocal behavior, which is consistent with previous findings in trust games (Berg et al., 1995; Camerer, 2003; Fehr and Gächter, 2000). However, of the 55% of subjects who demonstrate some form of reciprocal behavior, half of reciprocators (51%) discriminate by engaging at least partially in type-based reciprocity. As a

result, lab-based measures are likely overestimating the extent of dispositional reciprocity and fail to capture the many ways in which we treat individuals differentially based on perceptions formed outside the lab. While we do not know how these type-based reciprocators in our study would have behaved in an anonymous setting, one guess is that they would have simply taken the average amount of their type-based back transfers and given it to each subject, keeping their total back transfers fixed. While such a guess is speculative, we note that the overall average back transfer in our study is similar to the average back transfer in previous anonymous trust games.

To consider our findings in a broader perspective, we now turn to evolutionary explanations for human altruism and cooperation, as it may be easy to misinterpret our findings with respect to evolutionary "green beards;" in particular, one may be tempted to conclude that we have rejected a certain set of facial features from functioning as a "green beard" for trustworthiness. Such a conclusion is not warranted, but it is often asserted in the literature.

Let us first return to the three requirements for a "green beard" that could produce altruism or cooperation presented in Hamilton (1964): (i) an observable feature (phenotype), (ii) perception of the observable feature by others, and (iii) discrimination based on perception of the feature (Hamilton's words were, "social response consequent upon what was perceived"). On a superficial level, our findings support all three requirements; nonetheless, they cannot provide a basis for the evolution of altruism or cooperation because the feature itself is uninformative in our setting, so it fails to signal some underlying gene for altruism or cooperation.

A green beard would be effective, in the traditional evolutionary sense, if it was an observable *and* observed feature that results in one of two non-random aspects of interaction. First, interaction itself could be conditioned on possessing the green beard, leading to non-random matching (i.e. assortment), so that green beards interact with other green beards (and similarly for non-green beards). Second, under random matching, a green beard is a costless and accurate signal which serves as a coordination device in

conditional cooperation (i.e. green beards are cooperative when interacting with other green beards, and they are non-cooperative when interacting with non-green beards; non-green beards are always non-cooperative).

The subtle but critical issue is that our design – and the typical supposed tests for green beards in humans, such as Oosterhof and Todorov (2008) and Efferson and Vogt (2013) – fail these criteria of non-random interaction. The first criterion, assortment, fails trivially, due to the perfect stranger matching in our design. However, our second criterion, a coordination device for conditional cooperation, also fails because only one individual in the interaction is able to identify the (potential) green beard; the counterpart, the other person in the photograph, participated in a fully anonymous interaction with a random parter, and therefore could not behave differentially based on an observable feature. The traditional "green beard" hypothesis implicitly makes no prediction for behavior in such settings precisely because the unobservability of the green beard rules out its role as a coordination device (for non-random matching or conditional cooperation).

Results from our experiment and others with fully or partially anonymous interaction seem, however, to rule out a specific type of green beard. Gardner and West (2010) divide green beards into four categories. We list these briefly in Table B.10. Individuals identified in our study as dispositional reciprocity would seem to be an example of Gardner and West (2010)'s "helping, obligate" green beard; however, even this categorization is incorrect, as the unconditional helping behavior does not result in differential benefits for non-green beards and green beards. Therefore, even when restricting attention to this specific subtype, our design still does not allow us to test for an evolutionary green beard.[2] We again emphasize, however, that this shortcoming is not unique to our paper; instead, it is present in many papers claiming to test for green beards in the evolutionary biology and psychology literatures. The control treatment in Eckel and Petrie (2011) in which subjects always viewed a photograph without a payment represents a clean design that could test

[2]Generally speaking, a helping obligate green beard will be observationally equivalent to a helping facultative green beard under perfect assortment, since the counterfactual of interacting with a non-green beard is never observed; with random matching, as in our design, this potential confound is avoided.

Type	Behavior	Occurrence	Affects only
Helping, facultative	Help	Conditional	Other green beards
Helping, obligate	Help	Unconditional	Other green beards
Harming, facultative	Harm	Conditional	Non-green beards
Harming, obligate	Harm	Unconditional	Non-green beards

Table B.10: Green beard types specified by Gardner and West (2010). In both obligate types of green beard, the same action is always performed and the cost always incurred by the green beard; however, the action results in differential effects for non-green beards and other green beards. Helping green beards increase the inclusive fitness of other green beards; harming green beards decrease the inclusive fitness of non-green beards.

for facultative, helping green beards. We note, however, that our motivation was primarily in the formation, use, and accuracy of perceptions in incentivized settings; we devoted a large amount of attention to the evolutionary issues because of their prominence in the biological and social sciences, having particular relevance for economists with respect to altruism, cooperation, and social preferences. In our view, it appears that controlled economic experiments can only potentially identify "facultative" green beards (those who restrict helping or harming behavior conditionally on the type of counterpart), as the benefit to a fixed behavior – monetary payoffs – does not differ based on type of counterpart.

A final clarification is important. "Green beards" may appear to function as costly signals, but they are in fact costless. The green beard effect is generally considered implausible and lacking evolutionary stability because of the appearance of mutants who have a green beard but are not cooperative (or altruistic, trustworthy, etc). Costly signaling, in economics and biology, can be effective because the cost of the signal differs based on underlying type, so that a signal can sometimes serve as an honest revelation of type (Spence, 1973; Grafen, 1990). For a costly signalling approach to explain cooperation among unrelated humans, see Gintis et al. (2001). Confusion between "green beards" and costly signals can become even more problematic when using observable "tags" as the basis for cooperation (Riolo et al., 2001), as these tags are simply assortment mechanisms and can potentially be either green beards or costly signals. Tags can also present

difficulties as they may initially be meaningless but acquire meaning over time, such as through migration (Efferson et al., 2008).

B.9 Conclusion

In this paper, we have provided evidence supporting the formation and use of trustworthiness perceptions in experiments with monetary payoffs. These perceptions can be isolated to (and manipulated by) specific facial features, and these features primarily affect trustworthiness perceptions through their use in assessing the positive affect (eg. friendliness and happiness) of the person. However, we also found that these facial features are uninformative regarding behavior in a trust game, both for first-movers and second-movers. An open question is why people form strong "first impressions" that are not accurate? And why have we not learned over time that these impressions are uninformative?

Another open question is what, if anything, provides a basis for valid perceptions. Among the possibilities are be other static features such as symmetry (Zaatari and Trivers, 2007), dynamic "behavioral" green beards such as the expression of genuine emotions (Krumhuber et al., 2007), or costly signals such as clothing, hairstyle, and language (Riolo et al., 2001).

While we have helped shed some light on potential sources of stereotyping and discrimination, it is only a small flicker in a vast darkness. Certain forms of discrimination can help improve market outcomes, as in the case of costly signaling revealing unobservable type, but many other forms of discrimination – perhaps the majority – have no relevance for economic efficiency and lead to unequal outcomes with no justifiable basis. Future research should continue to identify the sources of immediate perceptions about the characteristics of others, whether and how those perceptions are used, and whether there is any evidence whatsoever that these perceptions are accurate.

B.10 References

Andreoni, J. (1989): "Giving with impure altruism: Applications to charity and Ricardian equivalence," *Journal of Political Economy*, 1447–1458.

——— (1990): "Impure altruism and donations to public goods: a theory of warm-glow giving," *Economic Journal*, 100, 464–477.

——— (1995): "Cooperation in public-goods experiments: kindness or confusion?" *American Economic Review*, 891–904.

Andreoni, J. and J. Miller (2002): "Giving according to GARP: An experimental test of the consistency of preferences for altruism," *Econometrica*, 70, 737–753.

Berg, J., J. Dickhaut, and K. McCabe (1995): "Trust, reciprocity, and social history," *Games and Economic Behavior*, 10, 122–142.

Bertrand, M., D. Chugh, and S. Mullainathan (2005): "Implicit discrimination," *American Economic Review*, 95, 94–98.

Bertrand, M. and S. Mullainathan (2004): "Are Emily and Brendan more employable than Latoya and Tyrone? Evidence on racial discrimination in the labor market from a large randomized experiment," *American Economic Review*, 94, 991–1013.

Bohnet, I. and R. Zeckhauser (2004): "Trust, risk and betrayal," *Journal of Economic Behavior and Organization*, 55, 467–484.

Bolton, G. E. and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90, 166–193.

Brewster, Z. W. and S. N. Rusche (2012): "Quantitative Evidence of the Continuing Significance of Race Tableside Racism in Full-Service Restaurants," *Journal of Black Studies*, 43, 359–384.

Camerer, C. (2003): *Behavioral game theory: Experiments in strategic interaction*, Princeton University Press.

Charness, G. and M. Rabin (2002): "Understanding social preferences with simple tests," *Quarterly Journal of Economics*, 117, 817–869.

Chugh, D. (2004): "Societal and managerial implications of implicit social cognition: Why milliseconds matter," *Social Justice Research*, 17, 203–222.

Conlin, M., M. Lynn, and T. O'Donoghue (2003): "The norm of restaurant tipping," *Journal of Economic Behavior and Organization*, 52, 297–321.

Dirks, D. and S. K. Rice (2004): ""Dining While Black" Tipping as Social Artifact," *Cornell Hotel and Restaurant Administration Quarterly*, 45, 30–47.

Eckel, C. C. and R. Petrie (2011): "Face value," *American Economic Review*, 101, 1497–1513.

Efferson, C., R. Lalive, and E. Fehr (2008): "The coevolution of cultural groups and ingroup favoritism," *Science*, 321, 1844–1849.

Efferson, C. and S. Vogt (2013): "Viewing men's faces does not lead to accurate predictions of trustworthiness," *Scientific reports*, 3.

Fehr, E. and U. Fischbacher (2005): "Altruists with green beards," *Analyse & Kritik*, 27, 73–84.

Fehr, E. and S. Gächter (2000): "Fairness and retaliation: The economics of reciprocity," *Journal of Economic Perspectives*, 14, 159–181.

Fehr, E. and K. M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817–868.

Fischbacher, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10, 171–178.

Frank, R. H. (1987): "If homo economicus could choose his own utility function, would he want one with a conscience?" *American Economic Review*, 593–604.

——— (1989): "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience? Reply," *American Economic Review*, 79, 594–596.

Gardner, A. and S. A. West (2010): "Greenbeards," *Evolution*, 64, 25–38.

Gintis, H., E. A. Smith, and S. Bowles (2001): "Costly signaling and cooperation," *Journal of Theoretical Biology*, 213, 103–119.

Goldin, C. and C. Rouse (2000): "Orchestrating Impartiality: The Impact of' Blind" Auditions on Female Musicians," *American Economic Review*, 90, 715–741.

Grafen, A. (1990): "Biological signals as handicaps," *Journal of Theoretical Biology*, 144, 517–546.

Greenwald, A. G., D. E. McGhee, and J. L. Schwartz (1998): "Measuring individual differences in implicit cognition: the implicit association test." *Journal of Personality and Social Psychology*, 74, 1464.

Greiner, B. (2004): "An Online Recruitment System for Economic Experiments," in *Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht 63, Göttingen: Ges. für Wiss. Datenverarbeitung*, ed. by K. Kremer and V. Macho, 79–93.

Hamilton, W. D. (1964): "The genetical evolution of social behaviour. II," *Journal of Theoretical Biology*, 7, 17–52.

Harrington, J. E. (1989): "If homo economicus could choose his own utility function, would he want one with a conscience? Comment," *American Economic Review*, 79, 588–593.

Henrich, J. (2004): "Cultural group selection, coevolutionary processes and large-scale cooperation," *Journal of Economic Behavior and Organization*, 53, 3–35.

Krumhuber, E., A. S. Manstead, D. Cosker, D. Marshall, P. L. Rosin, and A. Kappas (2007): "Facial dynamics as indicators of trustworthiness and cooperative behavior," *Emotion*, 7, 730–735.

Leider, S., M. M. Möbius, T. Rosenblat, and Q.-A. Do (2009): "Directed altruism and enforced reciprocity in social networks," *Quarterly Journal of Economics*, 124, 1815–1851.

Levine, D. K. (1998): "Modeling altruism and spitefulness in experiments," *Review of Economic Dynamics*, 1, 593–622.

Noll, E. D. and S. Arnold (2004): "Racial Differences in Restaurant Tipping Evidence from the Field," *Cornell Hotel and Restaurant Administration Quarterly*, 45, 23–29.

Nowak, M. A. (2006): "Five rules for the evolution of cooperation," *Science*, 314, 1560–1563.

Nowak, M. A. and K. Sigmund (1998): "Evolution of indirect reciprocity by image scoring," *Nature*, 393, 573–577.

Oosterhof, N. N. and A. Todorov (2008): "The functional basis of face evaluation," *Proceedings of the National Academy of Sciences*, 105, 11087–11092.

Panchanathan, K. and R. Boyd (2003): "A tale of two defectors: the importance of standing for evolution of indirect reciprocity," *Journal of Theoretical Biology*, 224, 115–126.

Riolo, R. L., M. D. Cohen, and R. Axelrod (2001): "Evolution of cooperation without reciprocity," *Nature*, 414, 441–443.

Sethi, R. and E. Somanathan (2003): "Understanding reciprocity," *Journal of Economic Behavior & Organization*, 50, 1–27.

Spence, M. (1973): "Job market signaling," *Quarterly Journal of Economics*, 87, 355–374.

Stirrat, M. and D. I. Perrett (2010): "Valid Facial Cues to Cooperation and Trust Male Facial Width and Trustworthiness," *Psychological Science*, 21, 349–354.

Tiddeman, B., M. Stirrat, and D. I. Perrett (2005): "Towards realism in facial image transformation: Results of a wavelet mrf method," in *Computer Graphics Forum*, Wiley Online Library, vol. 24, 449–456.

Tirole, J. (1988): *The Theory of Industrial Organization*, MIT press.

Trivers, R. L. (1971): "The evolution of reciprocal altruism," *Quarterly Review of Biology*, 46, 35–57.

Vogt, S., C. Efferson, and E. Fehr (2013): "Can we see inside? Predicting strategic behavior given limited information," *Evolution and Human Behavior*.

Zaatari, D. and R. Trivers (2007): "Fluctuating asymmetry and behavior in the ultimatum game in Jamaica," *Evolution and Human Behavior*, 28, 223–227.

B.11 Appendix: Experimental instructions

B.11.1 Main group (Original German version)

Anleitung

Vielen Dank für Ihre Teilnahme an der heutigen Studie!

Im Verlauf der Studie werden Sie zunächst an einigen Entscheidungssituationen teilnehmen, und im Anschluss einen langen Fragebogen ausfüllen.

Ihre Bezahlung hängt von Ihren eigenen Entscheidungen ab, sowie von den Entscheidungen einer Gruppe von Teilnehmern, die vor kurzem hier im Labor Ihre Entscheidungen getroffen haben.

Bitte lesen Sie den Text auf den folgenden Seiten gründlich durch, um folgende Dinge zu verstehen:

- Die Regeln
- Wie die Entscheidungssituation funktioniert
- Wer in den Entscheidungssituationen Ihr Gegenüber sein wird
- Wie Sie in Abhängigkeit Ihrer Entscheidungen bezahlt werden

Im Anschluss an die Anleitung wird es ein paar Kontrollfragen geben, die sicherstellen sollen, dass Sie diese vier Dinge gründlich verstanden haben.

<u>Allgemeine Regeln</u>
Bitte bleiben Sie während der gesamten Studie still an Ihrem Platz sitzen, und kommunizieren Sie mit niemandem. Sollten Sie Sich nicht an diese Regeln halten, müssen wir Sie bitten die Studie **ohne Bezahlung** abzubrechen, und Sie werden für die Teilnahme an zukünftigen Studien gesperrt.

Sollten Sie nach dem Lesen dieser Anleitung oder während der Studie noch Fragen haben, heben Sie bitte die Hand, und es wird ein Experimentleiter an Ihren Platz kommen, um Ihnen zu helfen.

Grundsätzliches

Sie werden an **31 Runden** einer Entscheidungssituation teilnehmen.

- Am Anfang der ersten Runde werden Sie **entweder der Rolle A oder der Rolle B** zugeteilt. Diese Rolle behalten Sie bis zum Ende der Studie bei.

- In **jeder der 31 Runden** bekommen Sie per Zufall **ein neues Gegenüber** zugeteilt, das die jeweils andere Rolle übernimmt.

- **Am Ende der Studie wird *nur eine der 31 Runden* nach dem Zufallsprinzip ausgewählt.**

- **Ihre Bezahlung resultiert aus Ihrer Entscheidung, sowie der Entscheidung Ihres Gegenübers in dieser einen Runde.**

Wer sind Ihre Gegenüber?

- **Ihre Gegenüber sind reale Personen, die aber heute nicht im Labor anwesend sind**. Alle 31 Ihrer heutigen Gegenüber haben vor kurzem hier im Labor Ihre Entscheidungen gefällt und diese sind von uns aufgezeichnet worden. Fuer diese Teilnehmer galten die gleichen Regeln, die auch für Sie heute gelten (Details weiter unten). Im Anschluss wurden Fotos dieser Teilnehmer gemacht, welche Sie heute waehrend des Experimentes zu sehen bekommen. Alle Teilnehmer, die Sie heute sehen werden, haben uns Ihr Einverständnis zur Verwendung der Fotos und Entscheidungen gegeben.

- Während Sie Ihre Entscheidung treffen, werden Sie auf der linken Bildschirmhälfte **ein Foto Ihres momentanen Gegenübers** sehen. Dieser Teilnehmer hat die Entscheidung getroffen, die Ihre Auszahlung in dieser Runde beeinflussen wird, falls diese Runde auszahlungsrelevant ist.

- Sie werden feststellen, dass wir die Fotos bearbeitet haben: z.B. haben wir die Bilder mit Hilfe einer schwarzen Umrandung so ausgeschnitten, dass Sie **nur das Gesicht** Ihres Gegenübers sehen werden.

- **Ihr Gegenüber hat bei der Aufzeichnung seiner/ihrer Entscheidungen *kein* Foto gesehen, so dass Sie für Ihr Gegenüber anonym sind.**

- Ihr Gegenüber wird, ebenso wie Sie, auf Basis Ihrer Entscheidungen heute bezahlt. **Wir lassen Ihrem Gegenüber seine/ihre Verdienste im Anschluss an die Studie in Bar zukommen.**

(bitte auf nächster Seite fortfahren)

2. Beschreibung der Entscheidungssituation

Im Folgenden finden Sie eine Beschreibung der Entscheidungssituation, die Sie heute 31x durchlaufen werden.

Teilnehmer A:

- Am Anfang der jeweiligen Runde erhalten A und B jeweils einen **Anfangsbetrag von CHF 12**.

- **Entscheidung:** Nach dem Erhalt des Anfangsbetrags muss Teilnehmer A entscheiden, wieviele seiner CHF 12 er an den Empfänger B senden möchte. Er kann **entweder CHF 0 oder CHF 10 senden**.

- Der Betrag, den Teilnehmer A schickt, **wird verdreifacht**, bevor er an Teilnehmer B weitergesendet wird.

- B hat bereits in der Vergangenheit entschieden, wieviel er / sie vom empfangenen Betrag an A zurück schicken möchte, *für den Fall, dass er etwas von A empfängt*.

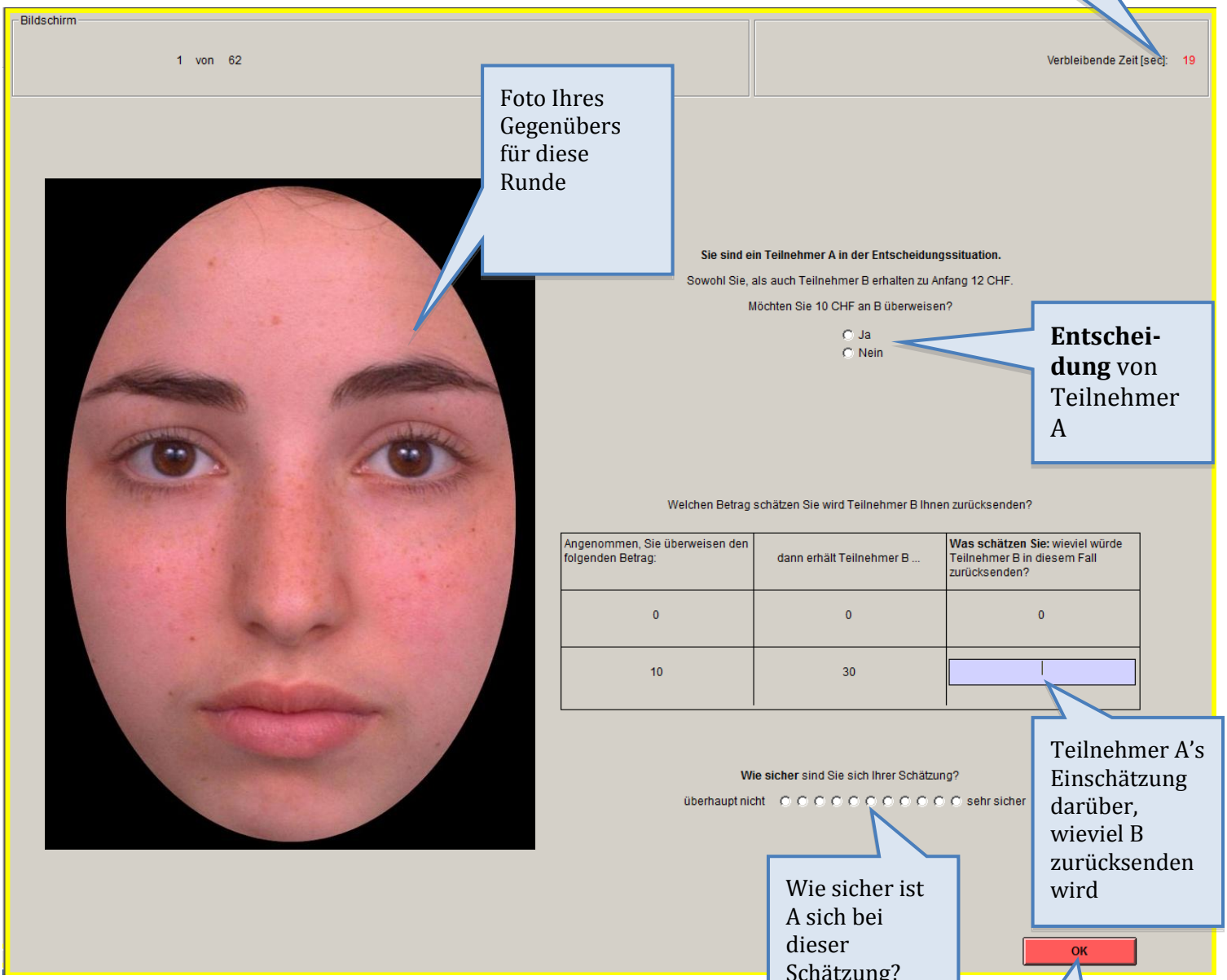
- Der **Verdienst** von Teilnehmer A für die gezogene Runde ist entsprechend:
  12 – (an B gesendeter Betrag) + (Betrag, den B zurücksendet)

- Teilnehmer A wird anschliessend um eine **Einschätzung** gebeten, wieviel er denkt, dass B zurücksenden würde, *für den Fall, dass er CHF 30 erhielte*. (Dies muss auch dann eingeschätzt werden, falls A nichts sendet.) Es wird ebenfalls gefragt, wie sicher A sich bei dieser Einschätzung ist.

Auf der nächsten Seite sehen Sie einen Screenshot des Entscheidungsbildschirms von Teilnehmer A.

(bitte auf nächster Seite fortfahren)

Entscheidungsbildschirm von Teilnehmer A:



(bitte auf nächster Seite fortfahren)

Teilnehmer B:

- Am Anfang der Runde erhält Teilnehmer B einen **Anfangsbetrag** von CHF 12.

- **Entscheidung:** Nach dem Erhalt des Anfangsbetrags muss B sich entscheiden, wieviel er an A zurücksenden möchte, *für den Fall, dass er etwas gesendet bekommt*. B muss diese Entscheidung fällen, *weiss dabei aber noch nicht, wieviel A gesendet hat.* Das bedeutet, B beantwortet die folgende Frage: „**Falls A CHF 10 sendet, so dass Sie 30 Franken bekommen: wieviel möchten Sie an A zurück senden?" (Zahl zwischen 0 und 30)**

- Falls sich später herausstellt, dass A nichts gesendet hat, dann erhalten beide Teilnehmer lediglich Ihr Startguthaben.

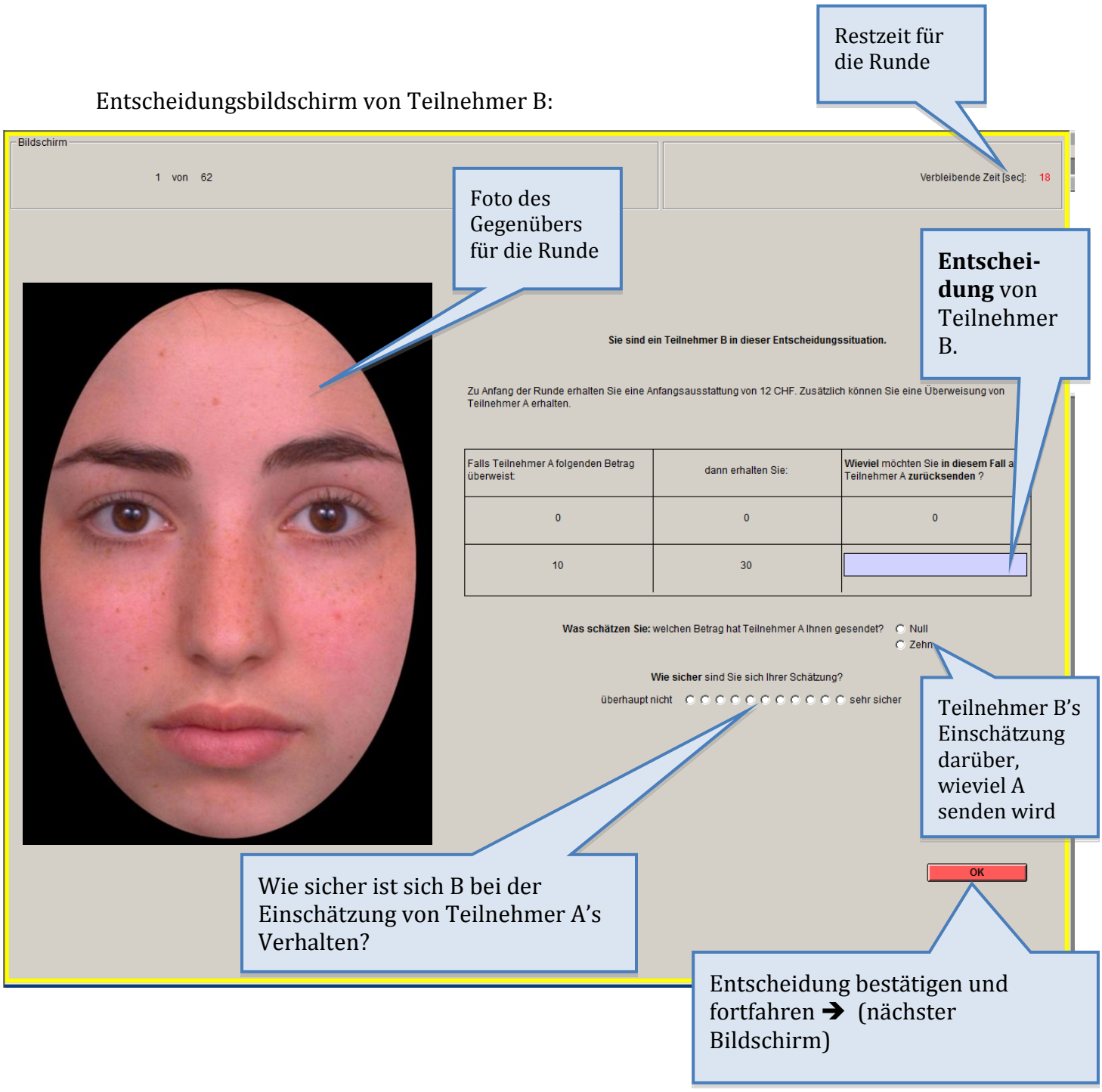
- Der Verdienst von Teilnehmer B für die gezogene Runde ist also:
  Falls A CHF 0 gesendet hat: 12 + 0 – 0
  Falls A CHF 10 gesendet hat: 12 + 30 – (zurück gesendeter Betrag)

- Teilnehmer B wird auch um eine **Einschätzung** dessen gebeten, ob er denkt, dass A zehn oder null Franken gesendet hat. Es wird ebenfalls gefragt, wie sicher B sich bei dieser Einschätzung ist.

Auf der nächsten Seite sehen Sie einen Screenshot des Entscheidungsbildschirms von Teilnehmer B.

(bitte auf nächster Seite fortfahren)

Entscheidungsbildschirm von Teilnehmer B:



(bitte auf nächster Seite fortfahren)

Die nächste Runde

Wenn Sie Ihre Entscheidung getroffen haben, endet die Runde und Sie kommen in die nächste Runde, mit einem neuen Gegenüber. Sie bekommen hierbei keine Rückmeldung darüber, wie sich Ihr Gegenüber in der vergangenen Runde entschieden hat.

Da Sie ausserdem nicht wissen können, welche der 31 Runden nach dem Zufallsprinzip für Ihre Zahlung relevant ist, sollten Sie jede Runde separat und für sich genommen betrachten.

Erst am Ende des Experimentes wird die auszahlungsrelevante Runde nach dem Zufallsprinzip bestimmt und Sie werden darüber informiert wie sich Ihr Gegenüber in dieser Runde entschieden hat. Sie erfahren aber nicht, um welche Runde es sich gehandelt hat. Das bedeutet, sie werden nicht erfahren, *welches* Ihrer Gegenüber diese Entscheidung getroffen hat. Dies stellt sicher, dass die Entscheidungen Ihrer Gegenüber anonym bleiben, obwohl Sie deren Gesichter sehen.

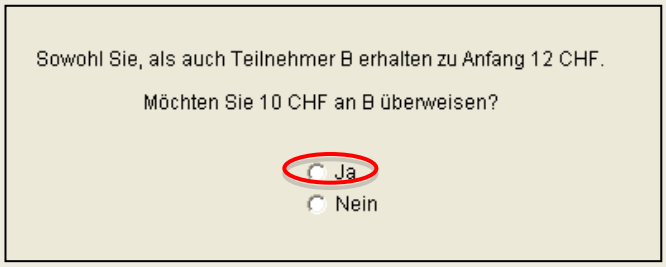
Bitte betrachten Sie das Beispiel auf der folgenden Seite, um ein Beispiel dafür zu sehen, wie die Auszahlungen funktionieren.

(bitte auf nächster Seite fortfahren)

Das folgende Beispiel illustriert, wie die Zahlungen bestimmt werden.

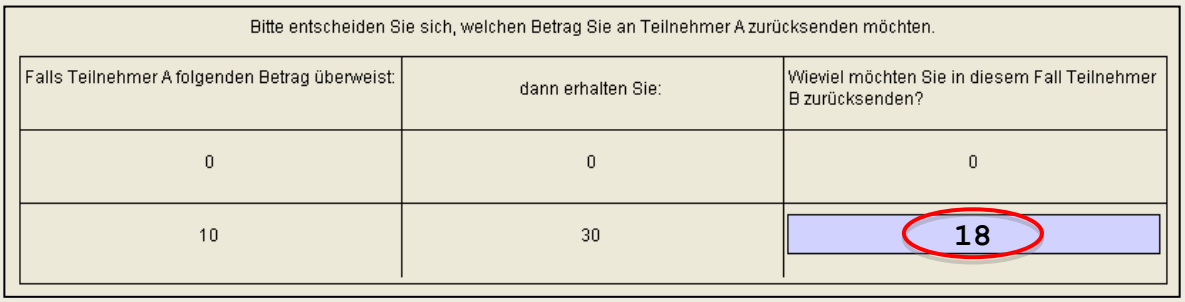
BEISPIEL

Am Anfang der Runde bekommt Teilnehmerin A 12 Franken.
Nehmen wir an, sie sendet 10 Franken an Teilnehmer B.



Der Betrag von 10 Franken wird nun verdreifacht, so dass Teilnehmer B 3x10=30 Franken erhalten wird.

Teilnehmer B weiss zum Zeitpunkt seiner eigenen Entscheidung allerdings noch nicht, ob er 0 oder 30 erhalten wird. Er bekommt ebenfalls 12 Franken am Anfang der Runde. Nehmen wir an, dass B sich entscheidet, 18 zurück zu senden, für den Fall, dass A den Betrag von 10 sendet (falls A null sendet, kann B nichts zurück senden):



Nun haben beide Teilnehmer ihre Entscheidungen getroffen. Falls diese Runde für die Auszahlung gezogen wird, ergeben sich die Auszahlungen wie oben beschrieben. Da Teilnehmer A 10 geschickt hat, wird die Entscheidung von B relevant, und dieser sendet 18 Franken zurück.

Also endet die Runde mit den folgenden Zahlungen:

Teilnehmer A:
12 – 10 + 18 = 20

Teilnehmer B:
12 + 30 – 18 = 24

(bitte auf nächster Seite fortfahren)

Fragebogen

Nachdem Sie die 31 Runden der Entscheidungssituation abgeschlossen haben, werden Sie gebeten, einen längeren Fragebogen auszufüllen.

Überblick: Zahlungen

- Unabhängig von Ihren heutigen Entscheidungen erhalten Sie in jedem Fall eine Teilnahmegebühr von CHF 10.
- Wie bereits erwähnt, wird eine der 31 gespielten Runden nach dem Zufallsprinzip gezogen, und Sie und Ihr Gegenüber erhalten die Zahlungen, die aus Ihren Entscheidungen in dieser Runde resultieren. Ihrem Gegenüber stellen wir die erwirtschafteten Beträge nach der Studie ebenfalls in Bar zu.

Wenn Sie fertig sind…

Wenn Sie die Studie abgeschlossen haben, **bleiben Sie bitte weiterhin still an Ihrem Platz sitzen und warten Sie darauf, dass der Experimentleiter Sie zur Auszahlung bittet.** Wenn Sie gerufen werden, bringen Sie bitte alle Ihre Sachen (Jacke, Tasche usw.) mit sich, da Sie das Labor durch den Auszahlungsraum verlassen werden.

…noch Fragen?

Falls Sie irgendwelche Fragen zu dieser Anleitung haben, heben Sie bitte nun die Hand, und ein Assistent wird an Ihren Platz kommen, um Ihnen zu helfen.

Kontrollfragen

Bitte füllen Sie nun den kurzen Verständnistest auf der nächsten Seite aus, um sicher zu gehen, dass Sie verstanden haben, wie die Auszahlungen der Entscheidungssituation sich zusammensetzen.

(bitte auf der nächsten Seite fortfahren)

B.11.2 Main group (English translation of earlier version)

INSTRUCTIONS

Thank you for participating in this study.

Today, you will engage in a number of interactions with a number of different people. Afterwards, you will be asked to complete a short survey. When the survey is complete, you will receive your payment in Swiss Francs and will be free to leave thereafter.

You will be paid based on your decisions, and on those of the other participants in the experiment.

Please carefully read the material on the following pages to understand

- The rules
- Who your counterparts will be
- How we will pay you

After you have read the instructions, **there will be a few test questions** to make sure you have understood these three things.

Please remain silent during the entire study, remain seated at your place, and refrain from communicating with anyone. Failure to comply with this may result in exclusion from the experiment without pay and removal from our list of participants for future experiments.

If you have any questions after reading these instructions or during the experiment, please raise your hand quietly and someone will come and assist you.

1. DESCRIPTION OF THE STUDY

There are two types of participants in this study: Proposers and Responders. You will be assigned one of these two roles. You will then take part in 30 rounds of a decision situation that is described below. In each round, you will be matched with a different counterpart.

Just like you, all of your counterparts have been fully informed about the rules and consequences of the decisions made in the game.

Note: In these instructions, we use "she" for proposers and "he" for responders for grammatical convenience only. You may be assigned to the role of either proposer or responder regardless of your gender. Similarly, your counterpart in any round may be male or female.

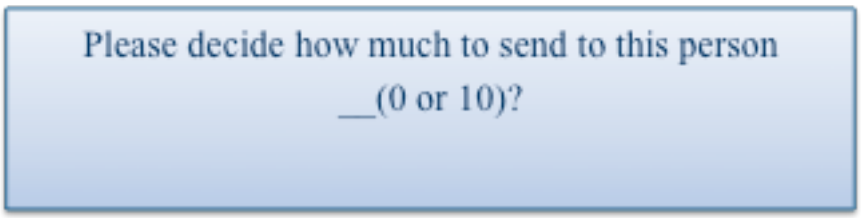
Initial endowment in each round

At the beginning of the round, both participants will receive an endowment of 12 tokens.

Proposers:

At the beginning of a round, the proposer decides how many of her 12 tokens that she wants to send to the responder. She may only send **either 0 or 10** tokens.

Thus, in each round, proposers answer the following question:

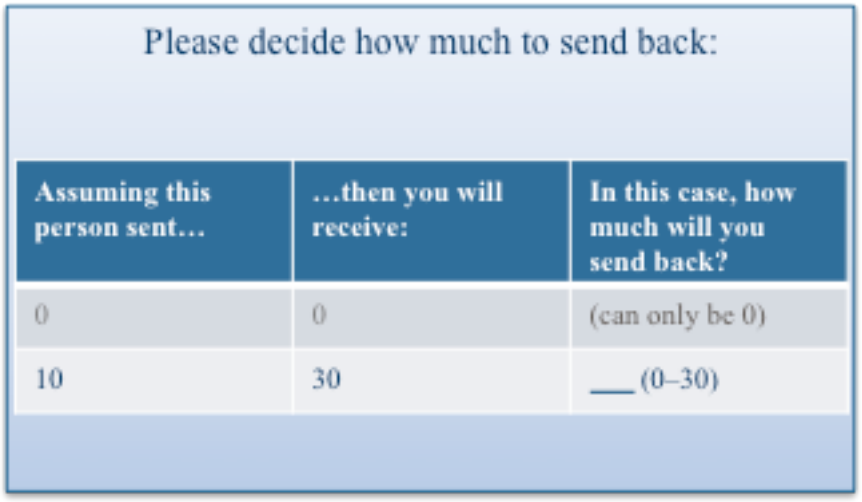


The amount the proposer sends **will be tripled** before being passed on to the responder. If the proposer chooses to send zero, then the round ends and both the proposer and the responder move on to the next round (with different partners). If the proposer sends 10 tokens, then the responder receives 30, and can decide how much of this to send back (any number between 0 and 30).

Responders

While the proposer decides how much to send, the responder has to decide how much to send back, in case he is going to receive 30 tokens. At this point, *the responder does not yet know whether the proposer is going to send anything*. If it turns out that the proposer sends nothing, then the responder's choice does not apply and nothing is sent back to the proposer.

That is, a responder answers the following question:



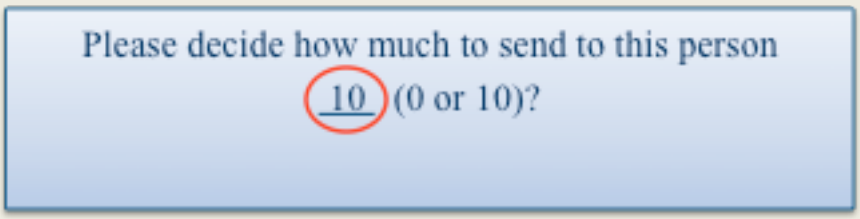
Once both participants have made their decisions, the round ends and both participants move on to the next round, with different partners.

At the end of the experiment, one of the rounds you played is randomly selected. The decisions in this round determine your payment, and that of your counterpart in that round.

The following example illustrates this.

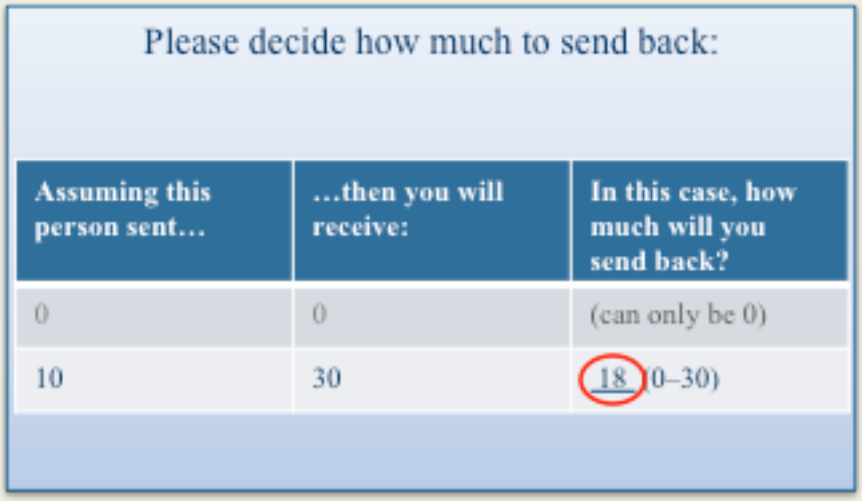
EXAMPLE

In the beginning, the Proposer is given 12 tokens.
Suppose she chooses to **send 10 tokens** to the Responder:



The number of tokens sent is then tripled, so the Responder will receive 3x10=30 tokens.
The responder, however, does not yet know how much the proposer sent.

The responder is also given 12 tokens in the beginning. Suppose he decides to **return 18**, *conditional on the sender sending 10* (if the proposer sends 0, then the responder has no choice but to return zero):



Now both members have made their choices, and payments are made according to the rules described earlier. Since the proposer sent 10 tokens, 18 tokens will be sent back by the responder.

Thus the session ends with the following payments:

Proposer's payment:
12 – 10 + 18 = 20

Responder's payment:
12 + 30 – 18 = 24

Your role

At the start of the study, you will be assigned to the role of either Proposer or Responder. You will maintain this role through the entire study.

Note that *you will NOT be given any information* about your counterparts' decisions until the end of the study. At the end of the study, **one of the trials will be randomly selected, and you will get paid based your decision and the decision of your counterpart**. Even though you will see the decision of the opponent that was randomly selected, and the payments you will receive, you will not see the picture of who this opponent was again (this is to maintain the anonymity of your counterparts).

Your counterpart

All of your counterparts today are people who participated in a similar study recently. On each trial, you will see a photograph of your counterpart on the left-hand side of the screen. You will notice that we have cut out the hair from the images so that only the face of the participant will be visible to you in front of a black background. **Your counterpart recorded his/her strategy in a recent study, and did not know whom they were going to be matched with today. Your counterpart did not see a photograph of you when they made their choices.**

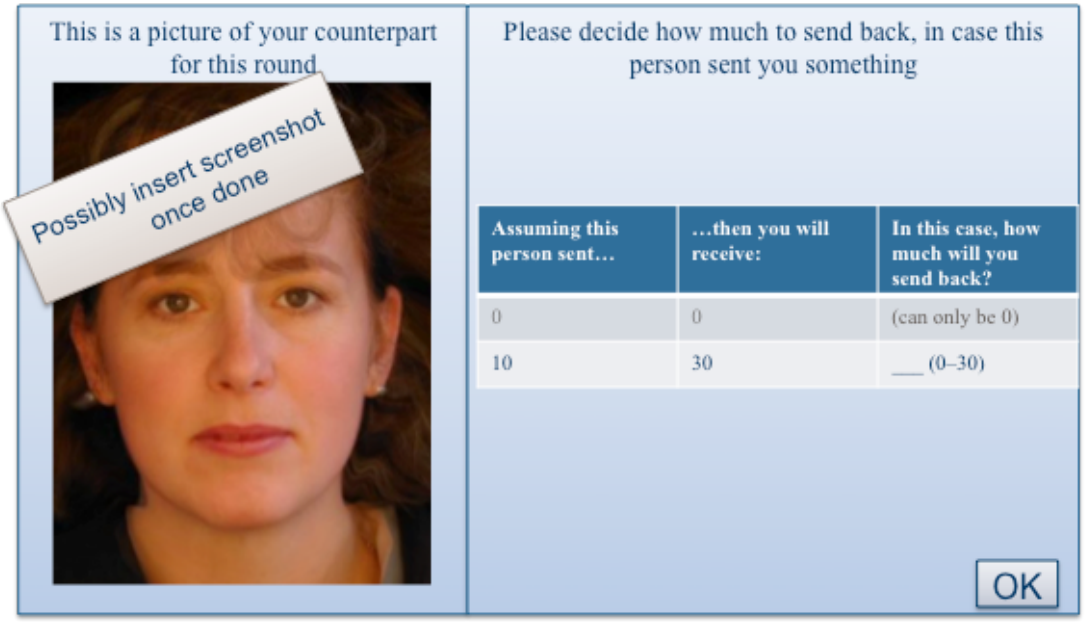
Your payment, and that of your counterpart, will depend on your decisions today, and on the pre-recorded decisions of your counterpart, in the round that is selected for payment. It is important to understand that ***even though your counterparts are not physically present today, your counterpart in the round selected for payment will also be paid based on your decisions and the rules described above.***

What your screen is going to look like

The following images are examples of the computer display that you will encounter, depending on whether you are a Proposer or a Responder.

If you are a proposer, you will see a screen that looks like this:

If you are a responder, you will see a screen that looks like the following:



SURVEY

After making all of your choices, you will then be asked to answer a short survey.

YOUR PAYMENT

At the very end of the study, one of your decisions will be selected at random, and you will be paid based on the decisions made by you and your counterpart on that trial. In addition, you will receive a show-up fee of 10 CHF.

Please consider each person you see separately and individually. This is important because only one of the trials will be drawn for payment, and you do not know in advance which one it is.

You will receive 1 CHF for each token that you earned. E.g. if you obtained 20 tokens on the trial selected for payment, your total payment would be 10 + 20 = 30 CHF.

WHAT TO DO WHEN YOU ARE DONE

When you are done, **please remain seated, continue to be silent and wait for the experimenter to call you into the payment room.** When you are called, please take all your belongings (bag etc.) with you, as you will exit through the payment room.

QUESTIONS?

If you have any questions about these instructions, please raise your hand now and the experimenter will help you.

COMPREHENSION TEST

On the next page, you will take a short comprehension test to ensure you understand the rules of the exchange and how these rules determine payments to Proposer and Responder.

B.11.3 Panel of raters (Original German version)

Anleitung

Vielen Dank für Ihre Teilnahme an der heutigen Studie!

Im Verlauf der heutigen Studie werden Sie die Gesichter von Personen sehen, die vor einigen Wochen an einer Studie hier im Labor teilgenommen haben. Ihre Aufgabe wird es sein, **zu erraten, wie die Personen auf den Bildern sich in der Studie verhalten haben**. Im Anschluss an diese Phase der Studie wird es einen Fragebogen geben.

Am Ende der Studie werden Sie in bar bezahlt. Zum einen erhalten Sie eine fixe Teilnahmegebühr iHv 20CHF. **Zum anderen erhalten Sie eine Bezahlung (bis zu 20CHF zusätzlich zur Teilnahmegebühr), deren Höhe sich danach richtet, wie präzise Sie das Verhalten der Personen auf den Fotos erraten haben.**

Bitte lesen Sie den Text auf den folgenden Seiten gründlich durch, um folgende Dinge zu verstehen:

- Die Regeln der heutigen Studie
- Wie die Entscheidungssituation funktioniert, an der die Personen in den Bildern teilgenommen haben
- Wie Sie für die Genauigkeit Ihrer Einschätzungen bezahlt werden

Im Anschluss an die Anleitung wird es **ein paar Kontrollfragen** geben, die sicherstellen sollen, dass Sie diese drei Dinge gründlich verstanden haben.

<u>Allgemeine Regeln</u>
Bitte bleiben Sie während der gesamten Studie still an Ihrem Platz sitzen, und kommunizieren Sie mit niemandem. Sollten Sie Sich nicht an diese Regeln halten, müssen wir Sie bitten die Studie **ohne Bezahlung** abzubrechen, und Sie werden für die Teilnahme an zukünftigen Studien gesperrt.

<u>Fragen?</u>
Sollten Sie nach dem Lesen dieser Anleitung oder während der Studie noch Fragen haben, heben Sie bitte leise die Hand, und es wird ein Experimentleiter an Ihren Platz kommen, um Ihnen zu helfen.

1. Wer sind die Personen auf den Fotos?

Vor ein paar Wochen haben wir eine Studie hier im Labor durchgeführt, bei der eine Gruppe von Personen an einer Entscheidungssituation teilgenommen hat. Im Anschluss wurden sie fotografiert.

- Alle Personen, die Sie heute sehen werden, haben uns ihre Erlaubnis erteilt, dass wir ihre Entscheidungen und Ihre Fotos in weiteren Studien in anonymer Form verwenden dürfen.
- Sie werden feststellen, dass wir die Fotos der Personen elektronisch bearbeitet haben -- wir haben eine ovale schwarze Schablone über das Gesicht gelegt, so dass die Haare nicht mehr sichtbar sind und nur das Gesicht sichtbar bleibt. Es handelt sich abgesehen hiervon um unveränderte Originalfotos dieser Personen.

2. Beschreibung der Entscheidungssituation, an der die Personen teilgenommen haben

- Die Personen haben mehrere Runden lang an einer Entscheidungssituation teilgenommen, die weiter unten beschrieben wird.
- Sie wurden dabei zu Anfangs jeweils einer von zwei Rollen zugeteilt: entweder Rolle A oder Rolle B. Diese Rolle wurde am Anfang festgelegt und bis zum Schluss beibehalten.
- In jeder Runde der Studie wurde eine Person der Rolle A jeweils einer Person der Rolle B zugeordnet. Wer wem zugeordnet wurde, wurde jede Runde geändert.
- Die Personen wussten dabei nicht, wer ihr Gegenüber ist – sie haben auch kein Foto ihres Gegenübers gesehen, sondern nur einen anonymen Entscheidungsbildschirm.
- In jeder Runde mussten die Teilnehmer eine Entscheidung fällen. Diese wird auf den folgenden Seiten beschrieben.

2a) Beschreibung der Entscheidungssituation von Teilnehmer A:

- Am Anfang der jeweiligen Runde erhalten A und B jeweils einen **Anfangsbetrag von 12 Schweizer Franken**.

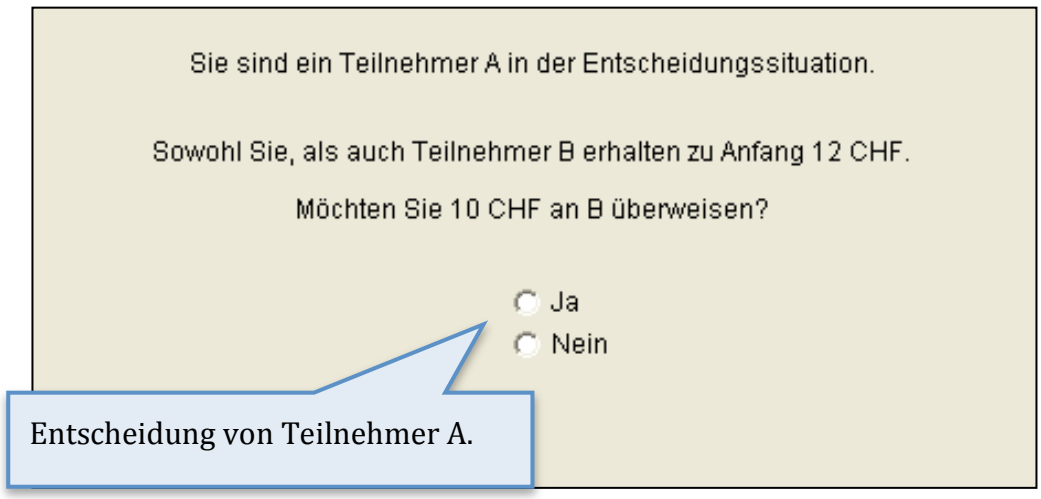
- **Entscheidung:** Nach dem Erhalt des Anfangsbetrags muss Teilnehmer A entscheiden, wieviele seiner 12 Franken er an den zugeteilten Teilnehmer B senden möchte. Er kann **entweder 0 oder 10 Franken senden**.

- Der Betrag, den Teilnehmer A schickt, **wird verdreifacht**, bevor er an Teilnehmer B weitergesendet wird. Bei B kommen also entweder 3*0=0 oder 3*10=30 Franken an.

- B entscheidet dann, *ohne zu wissen, ob A etwas gesendet hat*, wieviel er vom empfangenen Betrag an A zurück schicken möchte, *für den Fall, dass er etwas von A empfängt*.

- Der **Verdienst** von Teilnehmer A für die Runde ist:
  12 – (an B gesendeter Betrag) + (Betrag, den B zurücksendet)

Hier sehen Sie einen Screenshot des Entscheidungsbildschirms von Teilnehmer A:



(bitte auf nächster Seite fortfahren)

2b) Beschreibung Entscheidungssituation von Teilnehmer B:

- Am Anfang der Runde erhält Teilnehmer B einen **Anfangsbetrag** von 12 CHF.

- **Entscheidung:** Nach dem Erhalt des Anfangsbetrags muss B sich entscheiden, wieviel er an A zurücksenden möchte, *für den Fall, dass er etwas gesendet bekommt*.

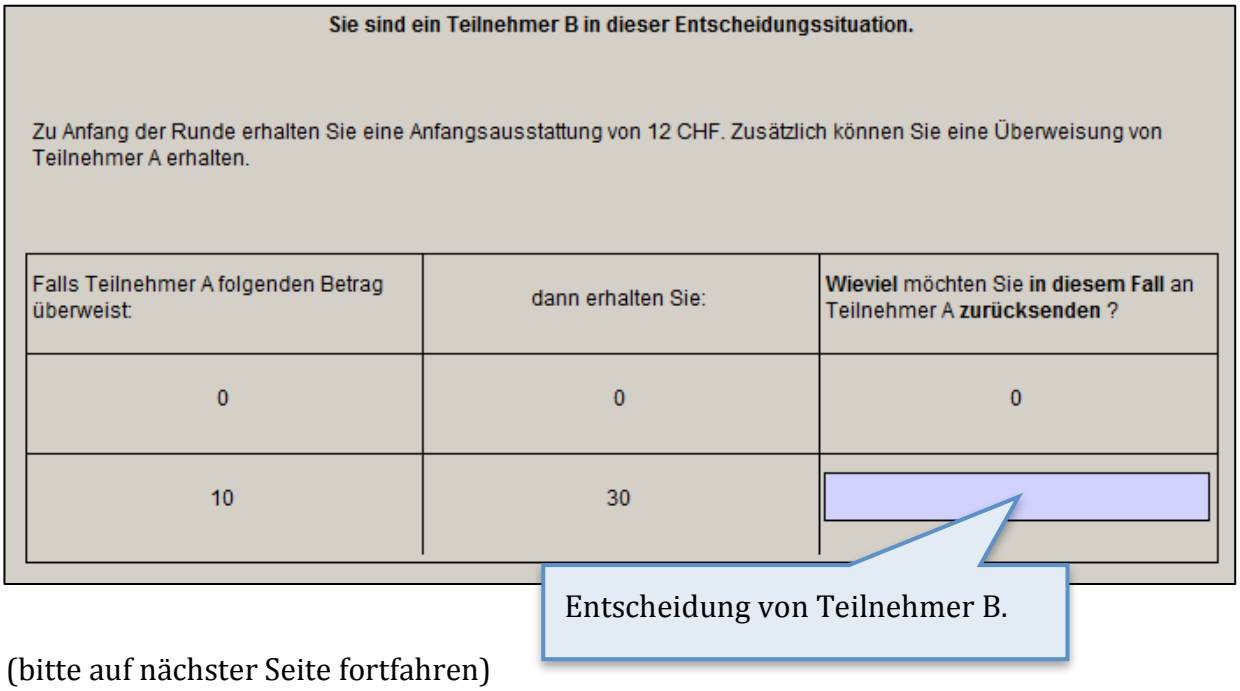
- B muss diese Entscheidung fällen, *weiss dabei aber noch nicht, wieviel A gesendet hat.* Das bedeutet, B beantwortet die folgende Frage: „**Falls A 10 Franken sendet, so dass Sie 30 Franken bekommen: wieviel möchten Sie an A zurück senden?“** B kann hierbei jeden Betrag zwischen (einschliesslich) 0 und 30 Franken zurückschicken.

- Falls sich später herausstellt, dass A nichts gesendet hat, dann ist die Entscheidung von B irrelevant, und es wird nichts an A zurück geschickt.

- Der Verdienst von Teilnehmer B ist also:
  Falls A null Franken gesendet hat: 12 + 0 – 0 = 12
  Falls A 10 Franken gesendet hat: 12 + 30 – (zurück gesendeter Betrag)

Hier sehen Sie einen Screenshot des Entscheidungsbildschirms von Teilnehmer B:

Sobald beide Teilnehmer Ihre Entscheidungen gefällt hatten, begann für die Teilnehmer die nächste Runde, mit einem neuen Gegenüber.

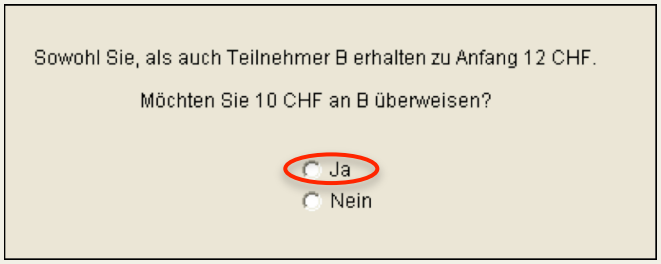


(bitte auf nächster Seite fortfahren)

Das folgende Beispiel illustriert, wie die Zahlungen von Teilnehmern A und B in der vergangenen Studie bestimmt wurden.

BEISPIEL

Am Anfang der Runde bekommt Teilnehmer A 12 Franken.
Nehmen wir an, er sendet 10 Franken an Teilnehmer B.



Der Betrag von 10 Franken wird nun verdreifacht, so dass Teilnehmer B 3x10=30 Franken erhalten wird.

Teilnehmer B weiss zum Zeitpunkt seiner eigenen Entscheidung allerdings noch nicht, ob er 0 oder 30 erhalten wird. Er bekommt ebenfalls 12 Franken am Anfang der Runde. Nehmen wir an, dass B sich entscheidet, 18 zurück zu senden, für den Fall, dass A den Betrag von 10 sendet (falls A null sendet, kann B nichts zurück senden):

Falls Teilnehmer A folgenden Betrag überweist:	dann erhalten Sie:	**Wieviel** möchten Sie **in diesem Fall** an Teilnehmer A **zurücksenden** ?
0	0	0
10	30	18

Nun haben beide Teilnehmer ihre Entscheidungen getroffen. Falls diese Runde für die Auszahlung gezogen wird, ergeben sich die Auszahlungen wie oben beschrieben. Da Teilnehmer A 10 geschickt hat, wird die Entscheidung von B relevant, und dieser sendet 18 Franken zurück.

Also endet die Runde mit den folgenden Zahlungen:

Teilnehmer A:
12 – 10 + 18 = 20

Teilnehmer B:
12 + 30 – 18 = 24

(bitte auf nächster Seite fortfahren)

3. Ihre Aufgabe und Ihre Zahlung, in Abhängigkeit der Genauigkeit Ihrer Schätzungen

Ihre Aufgabe ist es, **so genau wie möglich einzuschätzen, was die jeweilige Person auf dem Foto in der oben beschriebenen Entscheidungssituation getan hat**.

- Sie werden darum gebeten, für jeweils 30 Teilnehmer A und 30 Teilnehmer B solche Einschätzungen zu treffen.

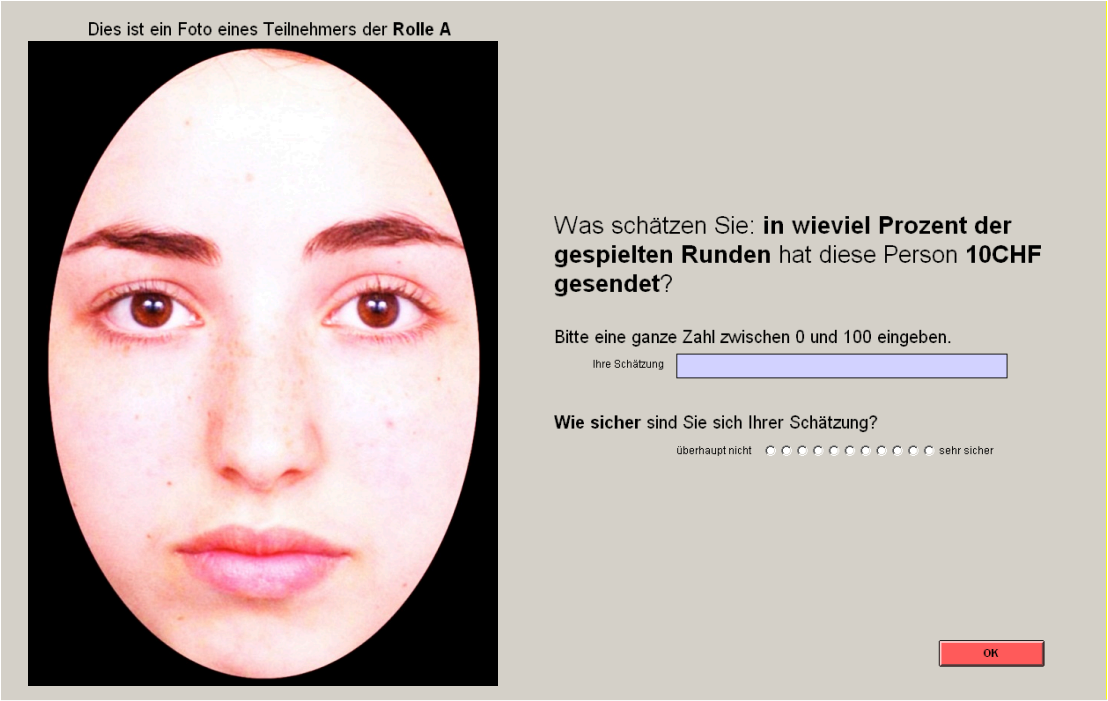
- Am Ende der Studie **wird nur eine dieser Personen zufällig ausgewählt**, und Sie werden danach bezahlt, wie präzise Ihre Einschätzung über die Entscheidungen dieser Person war. **Da nur eine Ihrer Einschätzungen ausbezahlt wird, sollten Sie jede Person separat und aufmerksam für sich genommen betrachten.**

- **Erst ganz zum Schluss der Studie**, nachdem Sie alle Ihre Einschätzungen getroffen haben und den Fragebogen ausgefüllt haben, werden Sie Rückmeldung darüber bekommen, wie präzise Ihre Schätzungen in der zufällig gezogenen Runde waren, und wieviel Sie dabei verdient haben. Sie werden allerdings *keine* Rückmeldung darüber bekommen, um *welche* Runde, bzw. welche dieser Personen es sich gehandelt hat, da wir die Anonymität der Entscheidungen unserer Teilnehmer A und B wahren müssen.

(bitte auf nächster Seite fortfahren)

3a) Einschätzungen bezüglich Teilnehmer A:

Falls die Person, die Sie sehen, ein Teilnehmer A war, wird Ihr Bildschirm wie folgt aussehen:



Wie Sie sehen, werden Sie gebeten, zwei Dinge einzugeben:

- In wieviel Prozent der Runden hat diese Person 0 vs. 10 an Teilnehmer B gesendet?
- Wie sicher sind Sie sich bei dieser Einschätzung?

Bezahlung für die Einschätzung von Teilnehmer A:

Ihre Bezahlung hängt davon ab, wie präzise Ihre Einschätzung ist. Sie werden mit einem Anfangsbetrag von 200 Punkten beginnen. Für jeden Prozentpunkt, mit dem Ihre Antwort von dem tatsächlichen Verhalten des Teilnehmers abweicht, bekommen Sie einen Punkt abgezogen:

Punkte = 200-2*(absoluter Fehlerbetrag A)

Das bedeutet, je näher Ihre Einschätzung an der korrekten Antwort ist, desto höher wird Ihre Bezahlung sein.

Sie bekommen zum Schluss die erzielten Punkte in Franken umgerechnet, mit dem Wechselkurs 1:10 (10 Punkte = 1 CHF)

Auf der folgenden Seite finden Sie ein Beispiel, um dies besser nachvollziehen zu können.

(bitte auf nächster Seite fortfahren)

BEISPIEL (Bezahlung für Einschätzung über einen Teilnehmer A)

Nehmen wir an, Sie hätten die Schätzungen in der zweiten Spalte abgegeben, und die korrekten Einschätzungen wären die in der dritten Spalte. Wie berechnet sich nun Ihre Bezahlung?

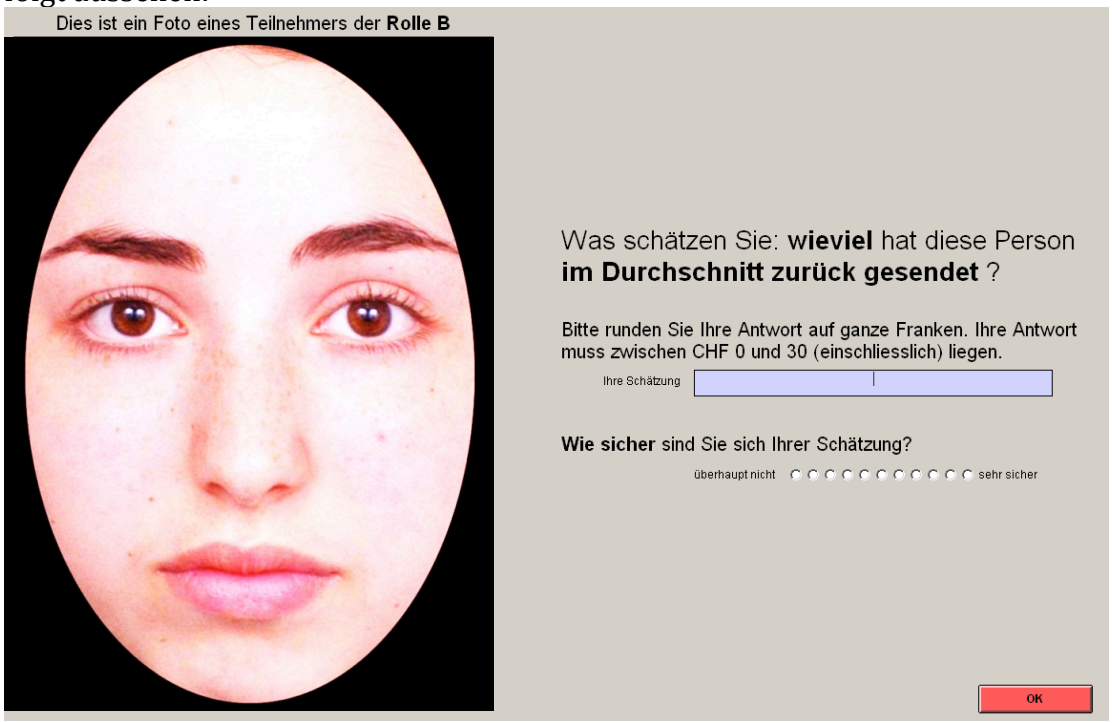
Ihre Schätzung: in wieviel % der Runden hat diese Person den Betrag 10 gesendet?	**Tatsächliches Verhalten der Person**	**Abweichung (Tatsächlich – Schätzung)**	**Absolutwert der Abweichung** *(Vorzeichen positiv machen)*
50%	40%	-10	10
			Abs. Fehlbetrag A=10

In diesem Beispiel ist der absolute Fehlerbetrag A = 10, also würden Sie **200-2*10=180** Punkte erhalten, sofern diese Runde zur Auszahlung gezogen wird.

Hierfür würden Sie also eine Bezahlung erhalten von 180/10=18,- CHF, welche auf Ihre Teilnahmegebühr von 20CHF aufgeschlagen würden. Sie würden in diesem Beispiel also insgesamt 38,- CHF verdienen.

(bitte auf nächster Seite fortfahren)

3b) Einschätzungen bezüglich Teilnehmern B:
Falls die Person, die Sie sehen, ein Teilnehmer B war, wird Ihr Bildschirm wie folgt aussehen:



Wie Sie sehen können, werden Sie für Teilnehmer B darum gebeten, eine Einschätzung zu geben für den *durchschnittlichen Betrag*, den die Person entschieden hat *zurück zu schicken, für den Fall, dass Teilnehmer A 10 sendet* (und somit 30CHF bei B ankommen).

Bezahlung für die Einschätzung von Teilnehmer B:
Auch hier gilt wieder, dass Ihre Bezahlung davon abhängt , wie präzise Ihre Einschätzung ist. Sie werden wiederum mit einem Anfangsbetrag von 200 Punkten beginnen. Für jeden Franken, den Sie daneben liegen, werden Ihnen fünf Punkte abgezogen:

Punkte = 200-5*(absoluter Fehlerbetrag B)

Das bedeutet, je näher Ihre Einschätzung an der korrekten Antwort ist, desto höher wird Ihre Bezahlung sein.

Sie bekommen zum Schluss die erzielten Punkte in Franken umgerechnet, mit dem Wechselkurs 1:10 (10 Punkte = 1 CHF)

Auf der folgenden Seite finden Sie ein Beispiel, um dies besser nachvollziehen zu können.

(bitte auf nächster Seite fortfahren)

BEISPIEL (Bezahlung für die Einschätzung von Teilnehmer Bs Verhalten)

Nehmen wir an, Sie hätten die Schätzungen in der zweiten Spalte abgegeben, und die tatsächlichen Entscheidungen seien in der dritten Spalte. Wie berechnet sich nun Ihre Bezahlung?

Ihre Schätzung: wieviel hat diese Person durchschnittlich zurückgeschickt, falls sie 3*10=30CHF erhalten hat?	**Tatsächliche Entscheidung der Person**	**Abweichung (Tatsächlich – Schätzung)**	**Absolutwert der Abweichung** *(Vorzeichen positiv machen)*
15	8	8-15=-7	7
			Abs. Fehlerbetrag B= 7

In diesem Beispiel ware **der absolute Fehlerbetrag B= 7, so dass Sie 200-7*5=165 Punkte erhalten würden, falls diese Runde für die Auszahlung gezogen wird.**

Dies entspräche 165/10=16,50 CHF, welche auf Ihre Teilnahmegebühr von 20CHF aufgeschlagen würden. Sie würden in diesem Beispiel also insgesamt 36,50CHF verdienen.

(bitte auf nächster Seite fortfahren)

4. Im Anschluss: ein Fragebogen…

Wenn Sie das Verhalten aller Teilnehmer A und Teilnehmer B geschätzt haben, werden Sie gebeten, am Bildschirm einen längeren Fragebogen auszufüllen.

5. Wenn Sie fertig sind…

Wenn Sie die Studie abgeschlossen haben, **bleiben Sie bitte weiterhin still an Ihrem Platz sitzen und warten Sie darauf, dass der Experimentleiter Sie zur Auszahlung bittet.** Wenn Sie gerufen werden, bringen Sie bitte alle Ihre Sachen (Jacke, Tasche usw.) mit sich, da Sie das Labor durch den Auszahlungsraum verlassen werden.

Überblick: Ihre Zahlungen

Um Ihnen einen zusammenfassenden Überblick zu geben: insgesamt werden Sie folgende Zahlungen erhalten:

1. In jedem Fall erhalten Sie mindestens eine Teilnahmegebühr von 20 CHF.

2. Zusätzlich erhalten Sie eine Bezahlung Ihrer Einschätzungen aus einer zufällig gezogenen Runde
    - Falls ein Teilnehmer A gezogen wird: *200-2*(abs. Fehlerbetrag A)*
    - Falls ein Teilnehmer B gezogen wird: *200-5*(abs. Fehlerbetrag B)*
    - Die Punkte werden mit einem *Wechselkurs* von 1:10 in Franken umgerechnet (10 Punkte = 1CHF). Falls Sie perfekt schätzen, bekommen Sie also 200/10=20CHF für Ihre Schätzung.
    - Um die Anonymität der Entscheidungen von Teilnehmern A und B zu wahren, bekommen Sie *keine Information, welche der Runden* zur Zahlung gezogen wurde.

…noch Fragen?

Falls Sie Fragen zu dieser Anleitung haben, heben Sie bitte nun die Hand, und ein Assistent wird an Ihren Platz kommen, um Ihnen zu helfen.

Verständnisfragen

Bitte füllen Sie nun den Verständnistest auf den nächsten Seiten aus, um sicher zu gehen, dass Sie verstanden haben, wie die Entscheidungssituation von Teilnehmern A und B aussah, und wie Ihre Bezahlung von Ihren Einschätzungen abhängt.

(bitte auf nächster Seite fortfahren)

B.11.4 Panel of raters (English translation of text)

INTRODUCTION

We are interested in using photographs of participants in future experiments, and your participation will assist in our endeavor. Researchers have identified a substantial number of ways that facial stimuli affect our perceptions and decisions. Unfortunately, many of these factors must be controlled for in experiments to identify causal influence. As a result, certain elements of these photographs must be removed. Our goal is to balance two conflicting goals: (i) remove elements of the photograph that will affect decisions in ways we cannot otherwise control for and (ii) maintain the integrity of the photographs, i.e. the photograph is believed to be of a real person and that this person is another participant in the current experiment.

For example, hairstyles can have strong effects, though we have little evidence to suggest how they have such effects. Since we do not know how hairstyles affect perceptions and decisions, we have removed hair from all of the photographs.

INSTRUCTIONS

For each of the [enter number here] photographs, you will be asked to rate the individual in the photograph on a variety of dimensions. In addition, you will be asked to make a prediction about the actual behavior of each of these people in an economic experiment.

DESCRIPTION OF THE GAME

The game consists of two players, Proposer and Responder. The game begins with the Proposer being in possession of a 12 tokens. The Proposer then chooses an amount of the tokens to the Responder; the Proposer may send 0, 4, 8, or 12 tokens. The number of tokens passed to the Responder is then tripled. Without knowing the Proposer's decision, the Responder is then able to pass any number tokens (from the tripled amount) back to the Proposer conditional on the amount sent.

EXAMPLE

The Proposer is given 12 tokens and chooses to send 8 tokens to the Responder. The number of tokens is then tripled, and the Responder receives 24 tokens. The Responder - without knowing how much the Proposer has sent - decides to send the following amounts back to the Proposer:

If the Proposer sent me 0 tokens:	send 0 tokens back to the Proposer
If the Proposer sent me 4 tokens:	send 2 tokens back to the Proposer
If the Proposer sent me 8 tokens:	send 9 tokens back to the Proposer
If the Proposer sent me 12 tokens:	send 14 tokens back to the Proposer

The game now ends with the following payoffs:

Proposer: 13 tokens (the 4 tokens kept from the initial allocation of 12 tokens plus the 9 tokens sent by the Responder)

Responder: 15 tokens (the 24 tokens received minus the 9 tokens sent back)

PAYMENT

You will be shown [enter number here - 20?] photographs and asked to rate them on a variety of dimensions. You will receive a fixed payment of [enter amount here - 25 CHF?] at the end of the session along with a performance-based payment. These payments are in addition to your earnings from the previous experiment. [NOTE: Assumes we're piggy-backing on another experiment.] We expect the total time for this session to be [enter time here - 30?] minutes.

PERFORMANCE-BASED PAYMENT

Each individual whose photograph you will see played the game 30 times in the role of either the Proposer or the Responder. You will be asked to predict the overall behavior for each individual.

For each Proposer, you will be asked the following:

(1) How many times (out of 30) did this individual send 0 tokens?

(2) How many times (out of 30) did this individual send 4 tokens?

(3) How many times (out of 30) did this individual send 8 tokens?

(4) How many times (out of 30) did this individual send 12 tokens?

Responders made 4 decisions in each of the 30 trials, for a total of 120 decisions. In each trial, Responders decided how much to send back to the Proposer for each possible amount sent by the Proposer. Thus, for each Responder, you will be asked the following:

(1) Over the 30 trials, what is the average number of tokens sent by this individaul given that the Proposer sent 0 tokens?

(2) Over the 30 trials, what is the average number of tokens sent by this individaul given that the Proposer sent 4 tokens?

(3) Over the 30 trials, what is the average number of tokens sent by this individaul given that the Proposer sent 8 tokens?

(4) Over the 30 trials, what is the average number of tokens sent by this individaul given that the Proposer sent 12 tokens?

At the end of the session, we will randomly select one Proposer and one Responder to calculate your performance-based payment.

PAYMENT - RESPONDER

error0 = | (your guess about the number of times the individual sent 0 tokens) - (actual number of times the individual sent 0 tokens) |

error4 = | (your guess about the number of times the individual sent 4 tokens) - (actual number of times the individual sent 4 tokens) |

error8 = | (your guess about the number of times the individual sent 8 tokens) - (actual number of times the individual sent 8 tokens) |

error12 = | (your guess about the number of times the individual sent 12 tokens) - (actual number of times the individual sent 12 tokens) |

total_error = error0 + error4 + error8 + error12

payment_proposer= 120 - total_error

error_avg0

error_avg4

error_avg0 = | (your guess about the average number of tokens sent by this individaul given that the Proposer sent 0 tokens) - (actual average number of tokens sent by this individaul given that the Proposer sent 0 tokens) |

error_avg4 = | (your guess about the average number of tokens sent by this individaul given that the Proposer sent 4 tokens) - (actual average number of tokens sent by this individaul given that the Proposer sent 4 tokens) |

error_avg8 = | (your guess about the average number of tokens sent by this individaul given that the Proposer sent 8 tokens) - (actual average number of tokens sent by this individaul given that the Proposer sent 8 tokens) |

error_avg12 = | (your guess about the average number of tokens sent by this individaul given that the Proposer sent 12 tokens) - (actual average number of tokens sent by this individaul given that the Proposer sent 12 tokens) |

total_avg_error = error_avg0 + error_avg4 + error_avg8 + error_avg12

payment_responder = 72 - total_avg_error

Your payment is: payment_proposer + payment_responder

EXAMPLE

We'll break this down into two sub-examples, one for Proposer and one for Responder.

SUB-EXAMPLE 1

Suppose you guessed that the Proposer in the photograph would send 0 tokens 5 times, 4 tokens 7 times, 8 tokens 3 times, and 12 tokens 15 times. This Proposer actually sent 0 tokens 6 times, 4 tokens 5 times, 8 tokens 9 times, and 12 tokens 10 times. Let's figure out the value of "payment_proposer."

error0 = | (your guess about the number of times the individual sent 0 tokens) - (actual number of times the individual sent 0 tokens) |

= | (5) - (6) |

= | -1 |

= 1

error4 = | (your guess about the number of times the individual sent 4 tokens) - (actual number of times the individual sent 4 tokens) |

= | (7) - (5) |

= | 2 |

= 2

error8 = | (your guess about the number of times the individual sent 8 tokens) - (actual number of times the individual sent 8 tokens) |

= | (3) - (9) |

= | -6 |

= 6

error12 = | (your guess about the number of times the individual sent 12 tokens) - (actual number of times the individual sent 12 tokens) |

= | (15) - (10) |

= | 5 |

= 5

total_error = error0 + error4 + error8 + error12

= 1 + 2 + 6 + 5

= 14

payment_proposer = 120 - total_error

= 120 - 14

= 106

SUB-EXAMPLE 2

Suppose you guessed that the Responder in the photograph would send an average of 0 tokens in response to a Proposer sending 0 tokens, an average of 2 tokens in response to a Proposer sending 4 tokens, an average of 17 tokens in response to a Proposer sending 8 tokens, and an average of 18 tokens in response to a Proposer sending 12 tokens.

This responder actually sent an average of 0 tokens in response to a Proposer sending 0 tokens, an average of 3 tokens in response to a Proposer sending 4 tokens, an average of 10 tokens in response to a Proposer sending 8 tokens, and an average of 9 tokens in response to a Proposer sending 12 tokens.

Let's figure out the value of "payment_responder."

error_avg0 = | (your guess given that the Proposer sent 0 tokens) - (actual average given that the Proposer sent 0 tokens) |

= | (0) - (0) |

= | 0 |

= 0

error_avg4 = | (your guess given that the Proposer sent 4 tokens) - (actual average given that the Proposer sent 4 tokens) |

= | (2) - (3) |

= | -1 |

= 1

error_avg8 = | (your guess given that the Proposer sent 8 tokens) - (actual average given that the Proposer sent 8 tokens) |

= | (17) - (10) |

= | 7 |

= 7

error_avg12 = | (your guess given that the Proposer sent 12 tokens) - (actual average given that the Proposer sent 12 tokens) |

= | (18) - (9) |

= | 9 |

= 9

total_avg_error = error_avg0 + error_avg4 + error_avg8 + error_avg12

= 0 + 1 + 7 + 9

= 17

payment_responder = 72 - total_avg_error

= 72 - 17

= 55

PAYMENT FROM SUB-EXAMPLES 1 AND 2

Your total payment is the sum of your payments from the two sub-examples. Let's figure out this amount.

Your payment is: payment_proposer + payment_responder

106 + 55

161

EXAMPLE

You will try another example on your own to make sure you understand the payment system.

[INSERT PRACTICE SCREEN HERE - You can use the following numbers:

Amount		Proposer (#times)	Guess	Error	Payment
0	5	10	5	100 (= 120 -20)	
4	9	12	3		
8	6	8	2		
12	10	0	10		

Amount		Responder (avg)	Guess	Error	Payment
0	0	0	0	59 (= 72 - 13)	
4	3	7	4		

8	10	12	2
12	24	17	7

END PRACTICE SCREEN]

Appendix C

Neural evidence for computational processes underlying risky decision making

This chapter is being prepared for submission to *Nature* and follows their formatting guidelines. Work in this chapter was conducted with Christopher J. Burke, Kerstin Preuschoff, Philippe N. Tobler, and Ernst Fehr. This chapter was written by Tony Williams.

C.1 Summary paragraph

Uncertainty pervades our daily lives and requires people to balance their preferences over outcomes with the likelihood of obtaining these outcomes. Two common approaches have been used to understand the neural decision-making processes in these situations. The first approach, common in animal learning and neuroscience, assumes that options are evaluated using mechanisms such as foraging and conditioning (Daw et al., 2006; Kolling et al., 2012; Lee et al., 2012; Niv et al., 2012; Schultz et al., 1997). The other approach, common in economics and psychology, assumes that preferences are known, stable, and can be modeled with expected utility or prospect theory (Fehr and Rangel, 2011; Fox and Poldrack, 2009; Mas-Collel et al., 1995; Kahneman and Tversky, 1979). Direct comparisons of these approaches, however, are lacking but critical for understanding the neural processes underlying decision-making. Here we show that commonly used models from both approaches explain neural activity in the same regions, which are either encoding or comparing subjective values of options. Furthermore, we show that neural activity is best explained by a model from animal learning, even in subjects whose observed behavioral choices are best explained by prospect theory (Kahneman and Tversky, 1979; Caraco et al., 1980; Stephens, 1986). Our results indicate that neural activity can be explained using behavioral models from either approach but that specific behavioral models may lead to erroneous inferences about the underlying neural processes. Recent theoretical models suggest that the discrepancy can potentially be resolved using summary statistics and reinforcement learning that can generate prospect theory-like behavior, and these models provide an evolutionary basis for risky decision-making rooted in ethology and foraging (Niv et al., 2012; Stephens, 1986; Shen et al., 2014).

C.2 Main text

Daily life requires us to make decisions on a regular basis, and we do not always know the outcomes with certainty. Therefore, we need some mechanism for learning and comparing

values (Daw et al., 2006; Kolling et al., 2012; Lee et al., 2012; Levy and Glimcher, 2012; Rushworth and Behrens, 2008). Traditionally, neuroscience has taken animal learning, foraging, and reinforcement as the basis for neural processes of risky decision making (Daw et al., 2006; Kolling et al., 2012; Niv et al., 2012; Schultz et al., 1997). The social sciences have generally focused on descriptive theories of choice which are increasingly used to model subjective valuation in the brain (Fehr and Rangel, 2011; Fox and Poldrack, 2009; Mas-Collel et al., 1995; Kahneman and Tversky, 1979; Lee, 2013). To date, however, we lack direct comparison of these approaches.

We studied choice behavior and neural activity in 29 healthy subjects while they made decisions in a two-alternative forced choice task with risky options (Fig. 1; see Supplementary Information). Following common procedures in economics and psychology, we did not provide feedback on the outcome of either option after the trial in order to rule out shifting or adaptive risk preferences during the task. While reinforcement plays a major role in the development of preferences, economic theory generally assumes stable preferences while psychologists and psychiatrists may regard risk-taking behavior as a dispositional trait.

We considered three commonly used models of risk preferences used in neuroscience, economics, and psychology (see Supplementary Information). Expected utility is the dominant model in economics and assumes that individuals are either risk averse, risk neutral, or risk seeking (Mas-Collel et al., 1995). Prospect theory is popular in psychology and behavioral economics and allows for preferences over risk to depend on a reference point, allowing risk aversion over gains and risk seeking over losses (Kahneman and Tversky, 1979). We also consider a mean-variance-skewness model in which preferences can be described based on simple summary statistics. The mean-variance approach has played a major role in optimal foraging theory and animal learning and also has a long history in finance (Caraco et al., 1980; Stephens, 1986; Markowitz, 1952). We add the third moment, skewness, to the model due to recent evidence suggesting neural encoding of skewness (Burke and Tobler, 2011; Symmonds et al., 2011; Wu et al., 2011).

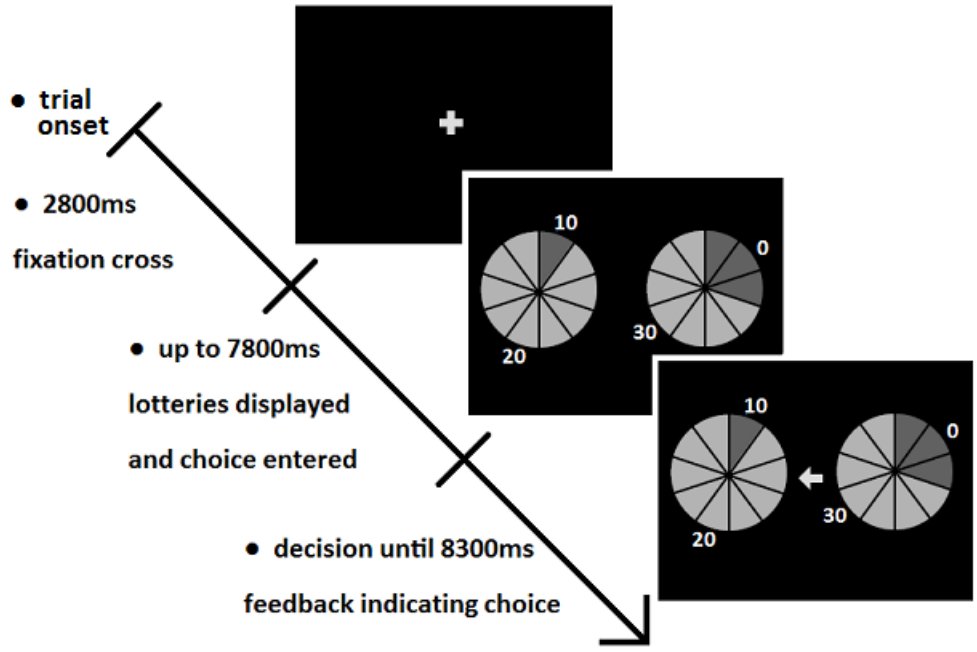


Figure C.1: **Task design.** At the start of each trial, a fixation cross is displayed for 2.8s. Lotteries are displayed and subjects allowed to enter choice for up to 5s (7.8s after trial onset). Feedback on lottery choice presented from time of decision until 5.5s after lotteries appear on screen (8.3s after trial onset). No additional feedback is given in the scanner. On each trial, subjects choose between one of the two options displayed on screen. Each pie slice denotes 10 percent probability. Number of points corresponding to the probability is outside the shaded area and indicates potential winnings.

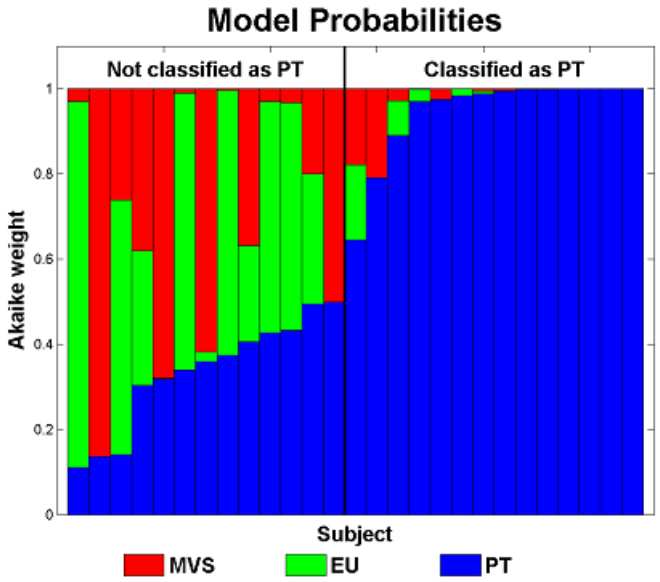


Figure C.2: **Model probabilities**. Individual subject model fits. Model probabilities are Akaike weights calculated using AICc, a small-sample corrected version of Akaike Information Criterion; 14 of 27 subjects are classified as prospect theory types at the behavioral level. For further details, see Tables C.9 and C.10.

We compared the fit of these three models for each individual subject. Models were fit using maximum likelihood and assumed a stochastic choice rule that allowed mistakes to happen more often when they had a low cost (see Supplementary Information). We compared the models using a small-sample corrected version of Akaike Information Criterion. The comparison allowed us to classify the behavior of 14 subjects as conforming to prospect theory, while we could not cleanly classify the remaining subjects as a single model type (Fig. 2, Supplementary Tables 8-10). Previous studies either assume that a single model is appropriate for all subjects without fitting other models or perform model fitting at the group level and select a single model for the population (Symmonds et al., 2011).

Given the parameter estimates from the behavioral models, we then calculated the subjective value for each risky option. We used these subjective values to create regressors containing the sum and the difference, in absolute value, of the subjective values for the options on each trial. The sum of subjective values should correlate with brain regions that encode value, while the difference in subjective values should correlate with brain

regions involved in value comparison. We used statistical parametric mapping to identify brain regions that significantly correlate with these model estimates ($p < 0.001$ voxelwise and $p_{FWE} < 0.05$ cluster-level corrected in each statistical parametric map). Regressors generated from the three behavioral models consistently identify the same brain regions. Value encoding, measured by the sum of subjective values, significantly correlates with right dorsolateral prefrontal cortex (Fig. 3A) and right lateral intraparietal cortex (Fig. 3B). Value comparison, measured by the difference in subjective values, significantly correlates with medial orbitofrontal cortex and ventromedial prefrontal cortex (Fig. 3C).

Since the behavioral models imply different computational processes, we wanted to see if the implied neural process based on the best-fitting behavioral model also best explains neural activity. To conduct this test, we used statistical parametric maps and recently developed methods from Rosa et al. (2010) to conduct Bayesian model selection on neural data (see Supplementary Information). We restrict attention to the 14 subjects who are classified as prospect theory types using behavioral choices. We further restrict attention to the neural models using prospect theory and mean-variance-skewness, as Bayesian analyses of functional neuroimaging data are computationally intensive and because the prospect theory and expected utility models gave very similar statistical results in our first analysis. Bayesian model selection provides strong evidence that the model with regressors containing mean-variance-skewness subjective values is substantially better at explaining the neural activity in right dorsolateral prefrontal cortex (Fig. 4A) and medial orbitofrontal cortex (Fig. 4B) than the model with regressors containing prospect theory subjective values. Importantly, this analysis only uses the subjects classified as prospect theory types at the behavioral level. The exceedance probability, the belief that the mean-variance-skewness model is better than the prospect theory model, is very high in both right dorsolateral prefrontal cortex ($\varphi = 0.98$, Fig. 4A) and medial orbitofrontal cortex ($\varphi = 0.96$, Fig. 4B).

These results are important for computational, behavioral, and neural accounts of decision making in risky environments as well as understanding neural development and

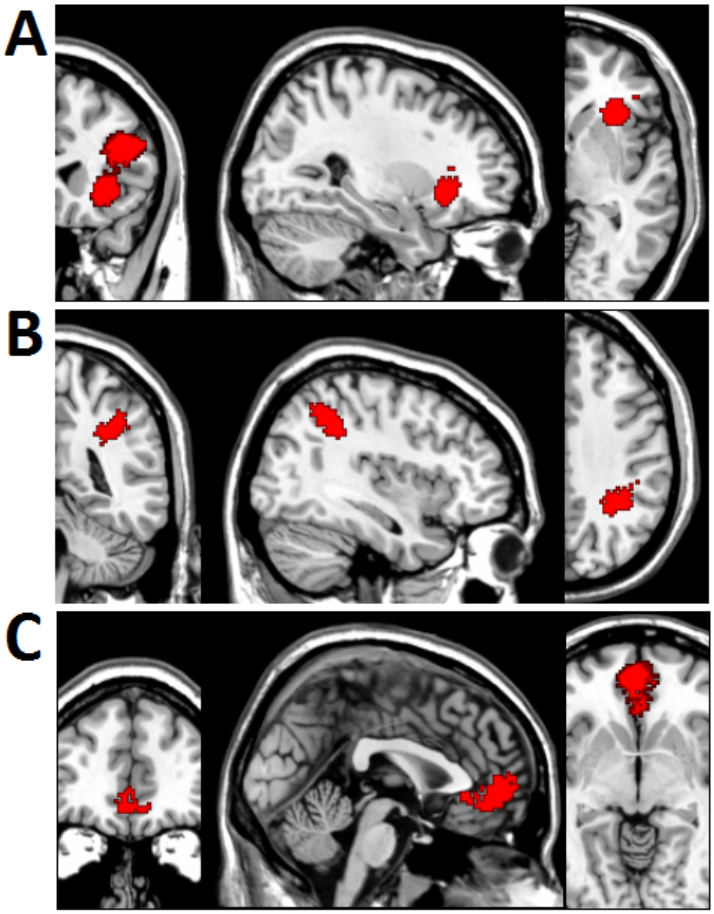


Figure C.3: **General subjective value encoding regardless of behavioral type.** Subjective value activations common to expected utility, prospect theory, and mean-variance-skewness computational models ($p < 0.0005$ whole-brain uncorrected and $p_{FWE} < 0.05$ cluster-level corrected in each contrast); images show intersection of contrasts. Value signal (sum of subjective values) correlated with activity in (A) right insula and dlPFC and (B) right LIP active over entire trial duration. Decision value (unsigned difference in subjective values) correlated with activity in (C) mOFC, vmPFC, and ACC at time of decision. Contrasts plotted at center of gravity for each cluster, defined as the average MNI x-,y-, and z-coordinates of the peak voxel from each computational model.

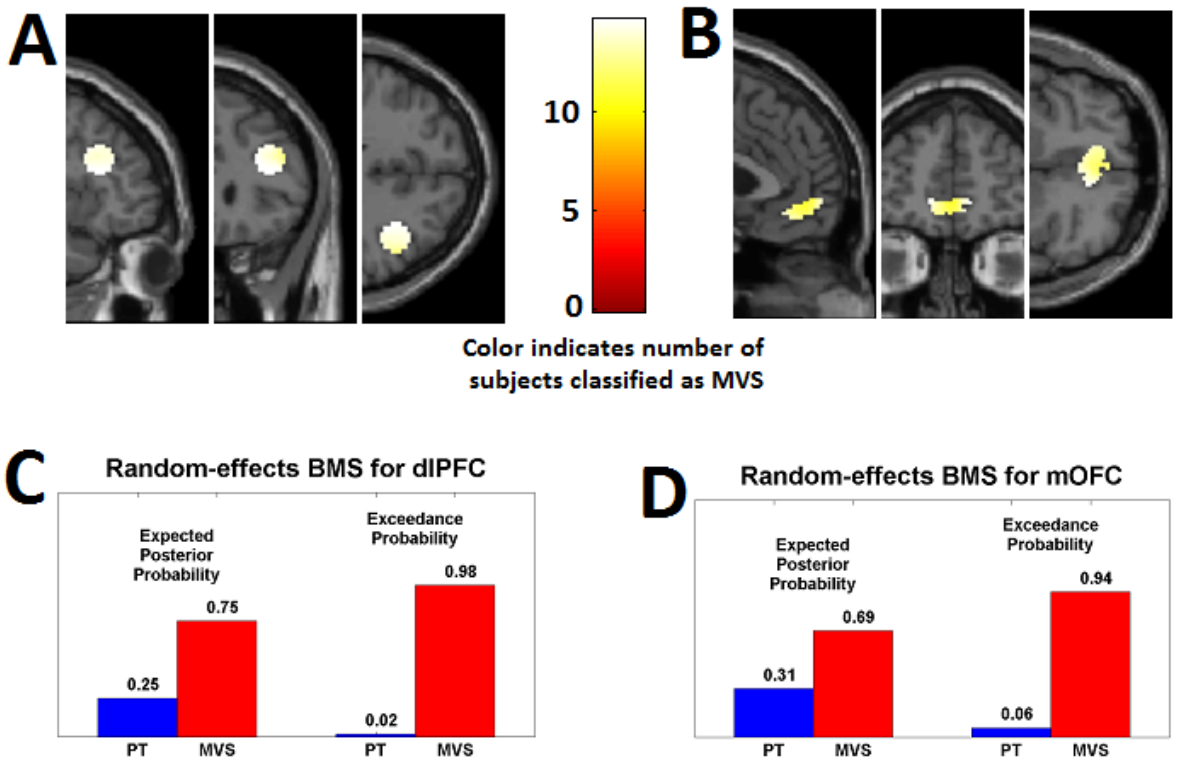


Figure C.4: **Behavioral prospect theory types are neural mean-variance-skewness types.** Random-effects Bayesian model selection at time of decision using only the 14 subjects classified as prospect theory (PT) types at the behavioral level. The α plots indicate the number of subjects for whom the mean-variance-skewness (MVS) model explains neural data better than PT model in (A) right dlPFC and (B) mOFC. Expected posterior probabilities and exceedance probabilities provide strong support in favor of MVS model in (C) right dlPFC and (D) mOFC.

maladaptation in psychiatric disorders. Behavioral models of risky decision making are intended, at most, to describe cognitive thought processes with no attention to neural functioning (Kahneman and Tversky, 1979; Mas-Collel et al., 1995). Computational and neural approaches have emphasized the need to forage in uncertain environments and learn values through prediction and reinforcement (Daw et al., 2006; Kolling et al., 2012; Lee et al., 2012; Niv et al., 2012; Schultz et al., 1997; Stephens, 1986). Several computational models suggest that the discrepancy between these approaches can potentially be reconciled. Risk-sensitive foraging models using only mean and variance can generate prospect theory-like behavior and suggest a metabolic reference point (Caraco et al., 1980; Stephens, 1986). A recent theoretical model uses risk-sensitivity and reinforcement learning to explicitly generate behavior consistent with prospect theory (Kahneman and Tversky, 1979; Shen et al., 2014). Our finding that subjective values encoding simple summary statistics best explains neural activity therefore does not contradict behavioral models of decision making but instead provides a neurobiological basis for such behavioral models as descriptive theories.

Given the prominent role of risk avoidance and risk seeking behavior in neural development and psychiatric disorders, we need a better understanding of the neural circuitry of risky decision making to understand maladaptations in development than can have major consequences (Blakemore and Robbins, 2012; Deisseroth, 2014; Selemon, 2013). Monoamine neurotransmitters commonly implicated in psychiatric disorders, such as serotonin, norepinephrine, and dopamine, have also been implicated in risky decision making (Schultz et al., 1997; Caplin et al., 2010; Doya, 2008; Juhasz et al., 2010; Kreek et al., 2005; Long et al., 2009). Recent interest has arisen in computational approaches to psychiatry both for diagnostic purposes and to identify potential treatments (Lee, 2013; Stephan and Mathys, 2014; Clark et al., 2013). Our results suggest that computational models from animal learning and foraging are potentially more informative for understanding neural processes of risky decision making than descriptive models borrowed from psychology and economics.

C.3 Methods

Thirty-three healthy human subjects participated in an fMRI experiment (Phillips Achieva 3T whole-body scanner) while making choices in a two-alternative forced choice task. Three subjects are excluded due to seeing only a subset of all trials; one subject is excluded due to a scanner crash. Choices were lotteries with earnings in points; four (of 180) choices were played out after the scanner task, and points were exchanged for money in Swiss francs. Three behavioral models were fit to choice data by maximizing the log-likelihood of the model given the parameters. Each behavioral model provided subjective values used as inputs to parametric regressors in two general linear models. General linear models were implemented using SPM8 (Wellcome Trust Centre for Neuroimaging, University College London). Two general linear models were estimated again using first-level Bayesian analysis and subjected to Bayesian model selection in SPM8. Further details are contained in Supplementary Information.

C.4 Supplementary Information

C.4.1 Supplementary Equations

Lotteries are denoted by x such that $x = (x_1, p_1; x_2, p_2)$, where p_i is the probability of obtaining outcome x_i, $x_1 \geq x_2$, and $p_1 + p_2 = 1$. A riskless lottery paying x_1 with certainty is denoted by $x = (x_1, 1; 0, 0)$.

Expected Utility Model

$$U(x) = p_1 x_1^\rho + p_2 x_2^\rho$$

in which we use a power utility function $u(x_i) = x_i^\rho$.

Prospect Theory Model

$$U(x) = w(p_1)x_1^\rho + [1 - w(p_1)]x_2^\rho$$

$$= x_2^\rho + w(p_1)[x_1^\rho - x_2^\rho]$$

$$= x_2^\rho + \left(\frac{\delta p_1^\gamma}{\delta p_1^\gamma + (1 - p_1)^\gamma}\right)[x_1^\rho - x_2^\rho]$$

in which we use the probability weighting function $w(p_i) = \delta p_i^\gamma / [\delta p_i^\gamma + (1 - p_i)^\gamma]$ and power value function $v(x_i) = x_i^\rho$. Due to the editing process in prospect theory, the smaller gain x_2 is perceived as a certain gain, and only the risky component $x_1 - x_2$ receives a probability weight, $w(p_1)$. Also note that we only use the nonnegative domain and two outcomes; therefore, the original and cumulative forms of prospect theory are equivalent to each other and to rank-dependent expected utility (Kahneman and Tversky, 1979; Quiggin, 1982; Tversky and Kahneman, 1992).

Mean-Variance-Skewness Model

$$U(x) = \beta_{\mu} Mean(x) + \beta_{\sigma} Variance(x) + \beta_{\gamma_1} Skewness(x)$$

Probabilistic Choice Rule

We use a Luce stochastic choice rule in which the probability of choosing lottery x when both x and y are available, $P(x|x,y)$ is given by

$$P(x|x,y) = \frac{e^{U(x)}}{e^{U(x)} + e^{U(y)}}$$

$$= \frac{1}{1 + e^{U(y)-U(x)}}.$$

Log-likelihood Function

$$\ln(L) = \sum_t \Big\{ \delta_t(x) P(x|x,y) + \big[1 - \delta_t(x)\big]\big[1 - P(x|x,y)\big] \Big\},$$

where $\delta_t(x)$ is a dummy variable equal to 1 (0) if lottery x (y) was selected on trial t.

C.4.2 Supplementary Methods

Experimental task

Subjects completed a risky decision-making task in the MRI scanner. At the beginning of each trial, two lotteries were displayed on the screen as pie charts; subjects had to choose which lottery they preferred. Subjects were instructed to make their decision by pressing the left or right button on a button box. Subjects were allowed up to 5s to make a decision; following the decision, feedback indicating the selected option (or failure to make a choice) was displayed until 5.5s after onset of trial. There were 180 trials consisting of lottery pairs evenly spaced among three functional runs. The lottery pairs are listed in Tables C.1–C.6. An additional 45 trials consisting of a prolonged fixation cross were interspersed between lottery trials and lasted for 5.5s. After completing the task in the scanner, three trials were selected for payment and played outside the scanner. Subjects earned points, which were converted to Swiss Francs (CHF) at the end of the experiment with 1 point = 0.25 CHF. In addition, subjects received 25 CHF for participating in the experiment. The experiment was conducted using Cogent 2000 developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience.

Analysis of behavioral data

Behavioral models were estimated using R statistical software (R Development Core Team, 2011). We obtained maximum likelihood estimates for the behavioral models using the `optimize` command for the one-variable expected utility model and the `optim` command and Nelder-Mead method for the prospect theory and mean-variance-skewness models by minimizing the negative log-likelihood (Nelder and Mead, 1965). For the expected utility and prospect theory models, we constrained the parameters to be positive by using the transformed variable $e^{\ln \psi}$ for each parameter ψ. For the mean-variance-skewness model, we did not constrain the parameter values. Since numerical optimiza-

tion procedures converge to local minima/maxima which do not necessarily correspond to the global minimum/maximum, we used several sets of starting values for parameters to increase the chances of finding the global minimum (i.e. maximum likelihood) parameter estimates. All sets of starting values for each model are listed in Table C.7. Parameter estimates for each subject are given in Table C.8. The log-likelihood, small-sample corrected Akaike Information Criterion (AICc), and Bayesian Information Criterion are listed for each model and subject in Table C.9. Given the model fits, we focus only on classifying subjects as prospect theory (PT) or unclassified, given lack of evidence supporting expected utility (EU) and mean-variance-skewness (MVS) types. Following conventional rules-of-thumb for model selection, a subject is classified as a PT type if (i) $AICc(EU) - AICc(PT) > 2$ and (ii) $AICc(MVS) - AICc(PT) > 2$ (Burnham and Anderson, 2002). Akaike weights and model classification for each subject are given in Table C.10

Analysis of fMRI data

Random effects whole brain analysis. We hypothesized that encoding of subjective values would occur from the appearance of lotteries on the screen until a decision was made and that comparison of subjective values would occur at the time of decision. These hypotheses suggested two distinct neural models. In the first neural model, the event occurs from trial onset until the time of decision, thus capturing neural activity related to value encoding. In the second neural model, the event is a stick function at time of decision, which should capture value comparison. We included parametric modulators for (i) the sum of subjective values for the two lotteries and (ii) the difference in subjective values in absolute value. We hypothesized that the sum of subjective values is a proxy for value encoding and would primarily correlate with neural activity in the first model. Similarly, the difference in subjective values should reflect value comparison and correlate with neural activity in the second model. We estimated both neural models using subjective values from each of the three behavioral models (expected utility, prospect

theory, and mean-variance-skewness) for a total of six distinct models (2 neural models $\times$ 3 behavioral models). Furthermore, we used a canonical haemodynamic response function (HRF) and included time and dispersion derivatives. Missed trials, in which a decision was not entered within 5s, and null trials, which consisted of only a fixation cross, were each modeled as separate conditions in the design matrix.

Bayesian Model Selection. Overall, we follow the procedures for Bayesian model selection in SPM described in Rosa et al. (2010). We repeated the same preprocessing as before but omitted the smoothing, as the log-evidence maps produced by the first-level analysis are smoothed before conducting the Bayesian model selection. We estimated two models using first-level Bayesian analysis. In each model, we included the sum of subjective values and the difference of subjective values (in absolute value) as parametric modulators with trials modeled using a stick function at the time of decision. The first model used subjective values from the mean-variance skewness model estimated from choice data. The second model used subjective values from the prospect theory model estimated from choice data. Log-evidence maps resulting from the first-level analysis were then smoothed using an 8mm full width at half maximum (FWHM) Gaussian kernel. We then conducted a voxel-wise random effects Bayesian model selection analysis. Given our interest in specific ROIs from our previous analysis, we created masks for MOFC and right DLPFC using PickAtlas. MOFC is defined using bilateral MOFC regions using anatomical AAL atlas. DLPFC is defined as an 8mm sphere around MNI coordinates [28 23 -2] with a dilation of 3; this voxel is the average of the peak voxel locations from the expected utility, prospect theory, and mean-variance-skewness models.

C.4.3 Supplementary Tables

Lottery Number	Left Lottery				Right Lottery			
	π_1	x_1	π_2	x_2	π_1	x_1	π_2	x_2
1	0.50	20	0.50	0	0.30	25	0.70	0
2	0.50	20	0.50	0	0.30	30	0.70	0
3	0.50	20	0.50	0	0.95	10	0.05	0
4	0.50	20	0.50	0	0.20	30	0.80	0
5	0.50	20	0.50	0	0.20	35	0.80	0
6	0.50	20	0.50	0	0.20	40	0.80	0
7	0.10	40	0.90	0	0.20	35	0.80	0
8	0.10	40	0.90	0	0.20	30	0.80	0
9	0.10	40	0.90	0	0.20	25	0.80	0
10	0.10	40	0.90	0	0.40	15	0.60	0
11	0.10	40	0.90	0	0.40	20	0.60	0
12	0.10	40	0.90	0	0.40	25	0.60	0
13	0.90	40	0.10	0	0.70	50	0.30	10
14	0.90	40	0.10	0	0.70	40	0.30	10
15	0.90	40	0.10	0	0.70	40	0.30	15
16	0.90	40	0.10	0	0.50	40	0.50	20
17	0.90	40	0.10	0	0.50	40	0.50	15
18	0.90	40	0.10	0	0.50	35	0.50	20
19	0.10	20	0.90	10	0.25	15	0.75	10
20	0.10	20	0.90	10	0.25	30	0.75	5
21	0.10	20	0.90	10	0.25	25	0.75	5
22	0.10	20	0.90	10	0.35	15	0.65	10
23	0.10	20	0.90	10	0.35	30	0.65	5
24	0.10	20	0.90	10	0.35	25	0.65	5
25	0.50	20	0.50	10	0.30	25	0.70	10
26	0.50	20	0.50	10	0.30	40	0.70	5
27	0.50	20	0.50	10	0.30	40	0.70	0
28	0.50	20	0.50	10	0.10	30	0.90	10
29	0.50	20	0.50	10	0.10	35	0.90	10
30	0.50	20	0.50	10	0.10	40	0.90	10

Table C.1: Lottery pairs 1–30

Lottery Number	Left Lottery				Right Lottery			
	π_1	x_1	π_2	x_2	π_1	x_1	π_2	x_2
31	0.90	20	0.10	10	0.80	25	0.20	5
32	0.90	20	0.10	10	0.80	30	0.20	5
33	0.90	20	0.10	10	0.70	30	0.30	5
34	0.90	20	0.10	10	0.70	35	0.30	0
35	0.90	20	0.10	10	0.60	35	0.40	5
36	0.90	20	0.10	10	0.60	35	0.40	0
37	0.05	40	0.95	10	0.20	20	0.80	10
38	0.05	40	0.95	10	0.20	15	0.80	10
39	0.05	40	0.95	10	0.30	20	0.70	10
40	0.05	40	0.95	10	0.30	25	0.70	10
41	0.05	40	0.95	10	0.40	20	0.60	10
42	0.05	40	0.95	10	0.40	25	0.60	10
43	0.25	20	0.75	10	0.40	25	0.60	5
44	0.25	20	0.75	10	0.40	40	0.60	0
45	0.25	20	0.75	10	0.50	25	0.50	5
46	0.25	20	0.75	10	0.15	45	0.85	5
47	0.25	20	0.75	10	0.15	40	0.85	5
48	0.25	20	0.75	10	0.05	45	0.95	10
49	0.75	20	0.25	10	0.60	20	0.40	15
50	0.75	20	0.25	10	0.60	30	0.40	5
51	0.75	20	0.25	10	0.50	25	0.50	10
52	0.75	20	0.25	10	0.50	30	0.50	5
53	0.75	20	0.25	10	0.90	20	0.10	5
54	0.75	20	0.25	10	0.90	25	0.10	0
55	0.95	40	0.05	10	0.75	40	0.25	25
56	0.95	40	0.05	10	0.75	35	0.25	25
57	0.95	40	0.05	10	0.60	50	0.40	20
58	0.95	40	0.05	10	0.60	45	0.40	20
59	0.95	40	0.05	10	0.50	45	0.50	25
60	0.95	40	0.05	10	0.50	35	0.50	30

Table C.2: Lottery pairs 31–60

Lottery Number	Left Lottery				Right Lottery			
	π_1	x_1	π_2	x_2	π_1	x_1	π_2	x_2
61	0.05	50	0.95	20	0.20	40	0.80	20
62	0.05	50	0.95	20	0.20	35	0.80	20
63	0.05	50	0.95	20	0.30	30	0.70	20
64	0.05	50	0.95	20	0.30	35	0.70	20
65	0.05	50	0.95	20	0.50	25	0.50	20
66	0.05	50	0.95	20	0.50	30	0.50	20
67	0.25	20	0.75	10	0.40	25	0.60	5
68	0.25	20	0.75	10	0.40	30	0.60	5
69	0.25	20	0.75	10	0.40	15	0.60	10
70	0.25	20	0.75	10	0.40	35	0.60	0
71	0.25	20	0.75	10	0.10	30	0.90	10
72	0.25	20	0.75	10	0.10	25	0.90	10
73	0.50	50	0.50	20	0.30	50	0.70	25
74	0.50	50	0.50	20	0.30	45	0.70	20
75	0.50	50	0.50	20	0.20	50	0.80	25
76	0.50	50	0.50	20	0.20	45	0.80	30
77	0.50	50	0.50	20	0.10	50	0.90	30
78	0.50	50	0.50	20	0.10	45	0.90	30
79	0.75	50	0.25	20	0.85	50	0.15	10
80	0.75	50	0.25	20	0.60	50	0.40	25
81	0.75	50	0.25	20	0.60	45	0.40	30
82	0.75	50	0.25	20	0.60	40	0.40	35
83	0.75	50	0.25	20	0.85	50	0.15	15
84	0.75	50	0.25	20	0.85	45	0.15	25
85	0.95	50	0.05	20	0.85	45	0.15	40
86	0.95	50	0.05	20	0.85	45	0.15	35
87	0.95	50	0.05	20	0.60	50	0.40	40
88	0.95	50	0.05	20	0.60	50	0.40	35
89	0.95	50	0.05	20	0.50	50	0.50	40
90	0.95	50	0.05	20	0.50	50	0.50	35

Table C.3: Lottery pairs 61–90

Lottery Number	Left Lottery				Right Lottery			
	π_1	x_1	π_2	x_2	π_1	x_1	π_2	x_2
91	0.5	20	0.5	0	1	13	0	0
92	0.5	20	0.5	0	1	10	0	0
93	0.5	20	0.5	0	1	7	0	0
94	0.1	40	0.9	0	1	5	0	0
95	0.1	40	0.9	0	1	7	0	0
96	0.1	40	0.9	0	1	10	0	0
97	0.9	40	0.1	0	1	25	0	0
98	0.9	40	0.1	0	1	30	0	0
99	0.9	40	0.1	0	1	33	0	0
100	0.1	20	0.9	10	1	7	0	0
101	0.1	20	0.9	10	1	10	0	0
102	0.1	20	0.9	10	1	15	0	0
103	0.5	20	0.5	10	1	12	0	0
104	0.5	20	0.5	10	1	15	0	0
105	0.5	20	0.5	10	1	17	0	0
106	0.9	20	0.1	10	1	15	0	0
107	0.9	20	0.1	10	1	17	0	0
108	0.9	20	0.1	10	1	19	0	0
109	0.05	40	0.95	10	1	12	0	0
110	0.05	40	0.95	10	1	15	0	0
111	0.05	40	0.95	10	1	18	0	0
112	0.25	20	0.75	10	1	12	0	0
113	0.25	20	0.75	10	1	15	0	0
114	0.25	20	0.75	10	1	17	0	0
115	0.75	20	0.25	10	1	15	0	0
116	0.75	20	0.25	10	1	17	0	0
117	0.75	20	0.25	10	1	19	0	0
118	0.95	40	0.05	10	1	35	0	0
119	0.95	40	0.05	10	1	33	0	0
120	0.95	40	0.05	10	1	30	0	0

Table C.4: Lottery pairs 91–120

Lottery Number	Left Lottery				Right Lottery			
	π_1	x_1	π_2	x_2	π_1	x_1	π_2	x_2
121	0.05	50	0.95	20	1	22	0	0
122	0.05	50	0.95	20	1	25	0	0
123	0.05	50	0.95	20	1	27	0	0
124	0.25	20	0.75	10	1	12	0	0
125	0.25	20	0.75	10	1	13	0	0
126	0.25	20	0.75	10	1	16	0	0
127	0.5	50	0.5	20	1	30	0	0
128	0.5	50	0.5	20	1	33	0	0
129	0.5	50	0.5	20	1	36	0	0
130	0.75	50	0.25	20	1	42	0	0
131	0.75	50	0.25	20	1	35	0	0
132	0.75	50	0.25	20	1	38	0	0
133	0.95	50	0.05	20	1	47	0	0
134	0.95	50	0.05	20	1	43	0	0
135	0.95	50	0.05	20	1	40	0	0
136	0.50	30	0.50	0	1.00	10	0.00	0
137	0.50	30	0.50	0	0.10	20	0.90	15
138	0.50	30	0.50	0	0.10	20	0.90	10
139	0.50	30	0.50	0	0.25	15	0.75	10
140	0.50	30	0.50	0	0.25	20	0.75	10
141	0.50	30	0.50	0	1.00	15	0.00	0
142	0.50	30	0.50	0	0.90	15	0.10	0
143	0.50	30	0.50	0	0.90	15	0.10	5
144	0.50	30	0.50	0	0.75	15	0.25	5
145	0.50	30	0.50	0	0.75	20	0.25	0
146	0.60	30	0.40	5	1.00	20	0.00	0
147	0.60	30	0.40	5	1.00	15	0.00	0
148	0.60	30	0.40	5	0.25	25	0.75	15
149	0.60	30	0.40	5	0.25	20	0.75	15
150	0.60	30	0.40	5	0.10	25	0.90	15

Table C.5: Lottery pairs 121–150

Lottery Number	Left Lottery				Right Lottery			
	π_1	x_1	π_2	x_2	π_1	x_1	π_2	x_2
151	0.40	30	0.60	5	1.00	15	0.00	0
152	0.40	30	0.60	5	1.00	10	0.00	0
153	0.40	30	0.60	5	0.75	20	0.25	10
154	0.40	30	0.60	5	0.75	20	0.25	5
155	0.40	30	0.60	5	0.90	20	0.10	5
156	0.50	50	0.50	0	1.00	20	0.00	0
157	0.50	50	0.50	0	0.10	30	0.90	15
158	0.50	50	0.50	0	0.10	30	0.90	20
159	0.50	50	0.50	0	0.25	30	0.75	15
160	0.50	50	0.50	0	0.25	30	0.75	20
161	0.50	50	0.50	0	1.00	15	0.00	0
162	0.50	50	0.50	0	0.90	25	0.10	10
163	0.50	50	0.50	0	0.90	20	0.10	15
164	0.50	50	0.50	0	0.75	25	0.25	10
165	0.50	50	0.50	0	0.75	20	0.25	15
166	0.60	50	0.40	10	1.00	30	0.00	0
167	0.60	50	0.40	10	0.10	45	0.90	25
168	0.60	50	0.40	10	0.10	40	0.90	25
169	0.60	50	0.40	10	0.25	40	0.75	25
170	0.60	50	0.40	10	0.25	35	0.75	25
171	0.40	50	0.60	10	1.00	25	0.00	0
172	0.40	50	0.60	10	0.90	25	0.10	15
173	0.40	50	0.60	10	0.90	30	0.10	15
174	0.40	50	0.60	10	0.75	25	0.25	15
175	0.40	50	0.60	10	0.75	30	0.25	15
176	0.50	40	0.50	10	1.00	25	0.00	0
177	0.50	40	0.50	10	1.00	20	0.00	0
178	0.50	40	0.50	10	0.25	30	0.75	20
179	0.50	40	0.50	10	0.75	25	0.25	15
180	0.50	40	0.50	10	0.10	30	0.90	20

Table C.6: Lottery pairs 151–180

	Expected Utility	Prospect Theory			Mean-variance-skewness		
	ρ	δ	γ	ρ	β_μ	β_σ	β_{γ_1}
1	0.25	0.25	0.25	0.25	0.01	-0.002	0.01
2	0.75	0.75	0.25	0.25	0.05	-0.002	0.01
3	1.00	1.00	0.25	0.25	0.10	-0.002	0.01
4		0.25	0.75	0.25	0.01	-0.001	0.01
5		0.75	0.75	0.25	0.05	-0.001	0.01
6		1.00	0.75	0.25	0.10	-0.001	0.01
7		0.25	1.00	0.25	0.01	0.000	0.01
8		0.75	1.00	0.25	0.05	0.000	0.01
9		1.00	1.00	0.25	0.10	0.000	0.01
10		0.25	0.25	0.75	0.01	-0.002	0.05
11		0.75	0.25	0.75	0.05	-0.002	0.05
12		1.00	0.25	0.75	0.10	-0.002	0.05
13		0.25	0.75	0.75	0.01	-0.001	0.05
14		0.75	0.75	0.75	0.05	-0.001	0.05
15		1.00	0.75	0.75	0.10	-0.001	0.05
16		0.25	1.00	0.75	0.01	0.000	0.05
17		0.75	1.00	0.75	0.05	0.000	0.05
18		1.00	1.00	0.75	0.10	0.000	0.05
19		0.25	0.25	1.00	0.01	-0.002	0.10
20		0.75	0.25	1.00	0.05	-0.002	0.10
21		1.00	0.25	1.00	0.10	-0.002	0.10
22		0.25	0.75	1.00	0.01	-0.001	0.10
23		0.75	0.75	1.00	0.05	-0.001	0.10
24		1.00	0.75	1.00	0.10	-0.001	0.10
25		0.25	1.00	1.00	0.01	0.000	0.10
26		0.75	1.00	1.00	0.05	0.000	0.10
27		1.00	1.00	1.00	0.10	0.000	0.10

Table C.7: Starting values for optimization routines.

Subject	Expected Utility	Prospect Theory			Mean-variance-skewness		
	ρ	δ	γ	ρ	β_μ	β_σ	β_{γ_1}
4	0.84	1.93	1.16	0.72	0.42	0.001	0.08
5	0.20	0.93	0.40	0.78	0.23	-0.002	0.40
6	0.36	0.70	0.53	0.58	0.09	-0.004	0.07
7	0.51	0.39	0.72	0.74	0.17	-0.009	0.00
8	0.20	0.87	0.20	0.77	0.01	0.000	0.45
9	0.89	1.44	0.90	0.82	0.43	0.002	0.01
10	0.81	0.87	0.87	0.87	0.48	-0.003	0.02
11	0.79	0.99	1.01	0.79	0.58	-0.005	0.20
13	0.64	2.64	1.68	0.53	0.19	-0.001	0.06
14	0.51	0.51	0.73	0.67	0.00	-0.004	-0.29
15	0.87	1.93	4.04	0.75	0.54	-0.001	0.07
16	0.70	0.77	0.83	0.79	0.44	-0.005	0.19
17	0.65	0.83	0.75	0.75	0.29	-0.004	0.07
18	0.84	1.22	0.79	0.83	0.34	0.002	-0.02
19	0.60	0.76	0.72	0.73	0.20	-0.003	0.03
20	0.49	0.59	0.60	0.70	0.22	-0.005	0.25
21	0.27	0.47	0.28	0.75	0.32	-0.011	0.85
22	0.80	0.73	0.92	0.90	0.74	-0.008	0.14
23	0.76	1.19	0.85	0.75	0.44	-0.002	0.23
24	0.46	0.31	0.64	0.68	0.19	-0.010	0.20
25	0.62	0.77	0.67	0.74	0.44	-0.006	0.42
26	0.75	0.72	0.88	0.86	0.35	-0.004	-0.22
27	0.68	1.80	1.63	0.55	0.27	-0.002	0.12
28	0.56	0.35	0.98	0.80	0.40	-0.016	0.11
29	0.48	0.25	0.78	0.78	0.15	-0.017	-0.04
30	0.57	0.59	0.81	0.69	0.25	-0.006	0.07
31	0.78	1.19	1.55	0.72	0.26	-0.001	-0.31

Table C.8: Individual parameter estimates for behavioral models.

Subject	N	Expected Utility			Prospect Theory			Mean-variance-skewness		
		ln(L)	AICc	BIC	ln(L)	AICc	BIC	ln(L)	AICc	BIC
4	172	-93.12	188.26	191.38	-84.14	174.42	183.72	-89.49	185.12	194.42
5	180	-121.12	244.27	247.44	-103.99	214.11	223.55	-115.87	237.88	247.32
6	178	-119.18	240.37	243.53	-115.82	237.78	247.19	-117.11	240.35	249.76
7	179	-106.56	215.13	218.30	-88.72	183.57	193.00	-98.21	202.55	211.98
8	177	-128.44	258.91	262.06	-98.05	202.24	211.63	-104.76	215.65	225.04
9	175	-91.63	185.29	188.43	-83.36	172.86	182.22	-83.36	172.85	182.21
10	178	-93.31	188.65	191.81	-92.70	191.54	200.94	-92.08	190.30	199.71
11	179	-90.38	182.78	185.94	-90.37	186.87	196.29	-91.67	189.47	198.90
13	179	-114.16	230.35	233.52	-105.40	216.93	226.36	-117.75	241.63	251.05
14	178	-108.88	219.78	222.94	-101.81	209.75	219.16	-106.91	219.96	229.36
15	174	-82.30	166.62	169.76	-67.26	140.67	150.00	-82.03	170.21	179.55
16	178	-101.43	204.88	208.04	-99.61	205.37	214.77	-102.28	210.70	220.11
17	167	-100.68	203.39	206.48	-99.13	204.40	213.61	-103.44	213.02	222.23
18	175	-100.33	202.68	205.82	-93.05	192.24	201.59	-92.29	190.72	200.07
19	169	-107.00	216.02	219.13	-105.59	217.32	226.56	-109.06	224.26	233.50
20	180	-112.39	226.80	229.98	-106.27	218.67	228.11	-113.91	233.96	243.41
21	178	-115.78	233.57	236.73	-83.47	173.08	182.48	-81.62	169.37	178.78
22	175	-81.84	165.70	168.84	-76.98	160.10	169.46	-76.44	159.02	168.37
23	180	-104.77	211.57	214.74	-102.24	210.61	220.06	-103.14	212.42	221.86
24	178	-107.30	216.61	219.77	-88.15	182.44	191.84	-97.42	200.97	210.38
25	170	-102.01	206.04	209.16	-99.99	206.12	215.39	-99.76	205.67	214.94
26	180	-93.36	188.75	191.92	-88.90	183.93	193.37	-92.32	190.78	200.22
27	175	-110.64	223.31	226.45	-108.80	223.73	233.08	-111.39	228.91	238.26
28	180	-104.30	210.62	213.80	-76.41	158.95	168.40	-77.73	161.61	171.05
29	174	-102.48	206.98	210.12	-69.47	145.09	154.42	-73.09	152.32	161.66
30	180	-105.44	212.91	216.08	-99.87	205.87	215.31	-106.18	218.50	227.94
31	158	-86.57	175.17	178.21	-83.93	174.02	183.05	-84.03	174.21	183.24

Table C.9: Log-likelihood and model selection values for behavioral models. Number of observations, N, varies by subject due to missed trials on which the subject did not enter a decision within 5s.

Subject	Expected Utility	Prospect Theory	Mean-variance-skewness	Classification
4	0.001	0.994	0.005	Prospect Theory
5	0.000	1.000	0.000	Prospect Theory
6	0.177	0.645	0.179	Prospect Theory
7	0.000	1.000	0.000	Prospect Theory
8	0.000	0.999	0.001	Prospect Theory
9	0.001	0.498	0.501	Unclassified
10	0.598	0.141	0.261	Unclassified
11	0.859	0.111	0.030	Unclassified
13	0.001	0.999	0.000	Prospect Theory
14	0.007	0.987	0.006	Prospect Theory
15	0.000	1.000	0.000	Prospect Theory
16	0.544	0.426	0.030	Unclassified
17	0.621	0.374	0.005	Unclassified
18	0.002	0.318	0.680	Unclassified
19	0.650	0.340	0.011	Unclassified
20	0.017	0.983	0.000	Prospect Theory
21	0.000	0.136	0.864	Unclassified
22	0.022	0.359	0.619	Unclassified
23	0.306	0.494	0.200	Unclassified
24	0.000	1.000	0.000	Prospect Theory
25	0.316	0.304	0.380	Unclassified
26	0.080	0.891	0.029	Prospect Theory
27	0.534	0.433	0.033	Unclassified
28	0.000	0.790	0.210	Prospect Theory
29	0.000	0.974	0.026	Prospect Theory
30	0.029	0.970	0.002	Prospect Theory
31	0.227	0.405	0.368	Unclassified

Table C.10: Akaike weights and model classification for individual subjects. Akaike weights calculated using AICc, a small-sample corrected version of AIC.

C.5 References

Blakemore, S.-J. and T. W. Robbins (2012): "Decision-making in the adolescent brain," *Nature neuroscience*, 15, 1184–1191.

Burke, C. J. and P. N. Tobler (2011): "Reward skewness coding in the insula independent of probability and loss," *Journal of Neurophysiology*, 106, 2415–2422.

Burnham, K. P. and D. R. Anderson (2002): *Model selection and multimodel inference: a practical information-theoretic approach*, Springer.

Caplin, A., M. Dean, P. W. Glimcher, and R. B. Rutledge (2010): "Measuring beliefs and rewards: a neuroeconomic approach," *The Quarterly Journal of Economics*, 125, 923–960.

Caraco, T., S. Martindale, and T. S. Whittam (1980): "An empirical demonstration of risk-sensitive foraging preferences," *Animal Behaviour*, 28, 820–830.

Clark, L., B. Averbeck, D. Payer, G. Sescousse, C. A. Winstanley, and G. Xue (2013): "Pathological Choice: The Neuroscience of Gambling and Gambling Addiction," *Journal of Neuroscience*, 33, 17617–17623.

Daw, N. D., J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan (2006): "Cortical substrates for exploratory decisions in humans," *Nature*, 441, 876–879.

Deisseroth, K. (2014): "Circuit dynamics of adaptive and maladaptive behaviour," *Nature*, 505, 309–317.

Doya, K. (2008): "Modulators of decision making," *Nature neuroscience*, 11, 410–416.

Fehr, E. and A. Rangel (2011): "Neuroeconomic foundations of economic choicerecent advances," *Journal of Economic Perspectives*, 25, 3–30.

Fox, C. R. and R. A. Poldrack (2009): "Prospect theory and the brain," in *Neuroeconomics: Decision making and the brain*, ed. by P. W. Glimcher, E. Fehr, C. Camerer, and R. A. Poldrack, Elsevier London, 145–173.

Juhasz, G., D. Downey, N. Hinvest, E. Thomas, D. Chase, Z. G. Toth, K. Lloyd-Williams, K. Mekli, H. Platt, A. Payton, et al. (2010): "Risk-taking behavior in a gambling task associated with variations in the tryptophan hydroxylase 2 gene: relevance to psychiatric disorders," *Neuropsychopharmacology*, 35, 1109–1119.

Kahneman, D. and A. Tversky (1979): "Prospect theory: An analysis of decision under risk," *Econometrica*, 263–291.

Kolling, N., T. E. Behrens, R. B. Mars, and M. F. Rushworth (2012): "Neural mechanisms of foraging," *Science*, 336, 95–98.

Kreek, M. J., D. A. Nielsen, E. R. Butelman, and K. S. LaForge (2005): "Genetic influences on impulsivity, risk taking, stress responsivity and vulnerability to drug abuse and addiction," *Nature neuroscience*, 8, 1450–1457.

Lee, D. (2013): "Decision making: from neuroscience to psychiatry," *Neuron*, 78, 233–248.

Lee, D., H. Seo, and M. W. Jung (2012): "Neural Basis of Reinforcement Learning and Decision Making," *Annual Review of Neuroscience*, 35, 287–308.

Levy, D. J. and P. W. Glimcher (2011): "Comparing apples and oranges: using reward-specific and reward-general subjective value representation in the brain," *Journal of Neuroscience*, 31, 14693–14707.

——— (2012): "The root of all value: a neural common currency for choice," *Current opinion in neurobiology*, 22, 1027–1038.

Long, A. B., C. M. Kuhn, and M. L. Platt (2009): "Serotonin shapes risky decision making in monkeys," *Social cognitive and affective neuroscience*, 4, 346–356.

Markowitz, H. (1952): "Portfolio selection," *Journal of Finance*, 7, 77–91.

Mas-Collel, A., M. Whinston, and J. Green (1995): "Microeconomic theory," .

Nelder, J. A. and R. Mead (1965): "A simplex method for function minimization," *Computer Journal*, 7, 308–313.

Niv, Y., J. A. Edlund, P. Dayan, and J. P. O'Doherty (2012): "Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain," *Journal of Neuroscience*, 32, 551–562.

Quiggin, J. (1982): "A theory of anticipated utility," *Journal of Economic Behavior & Organization*, 3, 323–343.

R Development Core Team (2011): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rosa, M. J., S. Bestmann, L. Harrison, and W. Penny (2010): "Bayesian model selection maps for group studies," *Neuroimage*, 49, 217–224.

Rushworth, M. F. and T. E. Behrens (2008): "Choice, uncertainty and value in prefrontal and cingulate cortex," *Nature neuroscience*, 11, 389–397.

Schultz, W., P. Dayan, and P. R. Montague (1997): "A neural substrate of prediction and reward," *Science*, 275, 1593–1599.

Selemon, L. (2013): "A role for synaptic plasticity in the adolescent development of executive function," *Translational psychiatry*, 3, e238.

Shen, Y., M. J. Tobia, T. Sommer, and K. Obermayer (2014): "Risk-sensitive Reinforcement Learning," *Neural Computation*, in press.

Stephan, K. E. and C. Mathys (2014): "Computational approaches to psychiatry," *Current Opinion in Neurobiology*, 25, 85–92.

——— (1986): *Foraging theory*, Princeton University Press.

Symmonds, M., N. D. Wright, D. R. Bach, and R. J. Dolan (2011): "Deconstructing risk: separable encoding of variance and skewness in the brain," *Neuroimage*, 58, 1139–1149.

Tversky, A. and D. Kahneman (1992): "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and Uncertainty*, 5, 297–323.

Wu, C. C., P. Bossaerts, and B. Knutson (2011): "The affective impact of financial skewness on neural activity and choice," *PloS ONE*, 6, e16838.

Appendix D

Curriculum Vitae

CURRICULUM VITAE

Tony Brett Williams

PERSONAL DETAILS

Date of Birth:	December 6, 1981
Place of Birth:	Maryland, United States of America
Citizenship:	United States of America

EDUCATION

09/2009–07/2014	Doctoral studies in Economics, University of Zurich
10/2008–10/2010	MSc in Social and Cultural Psychology, London School of Economics
09/2006–07/2008	MA in Economics, Johns Hopkins University
08/2004–08/2006	BS in Mathematics and Economics, Florida State University
08/1999–05/2003	BS in Political Science and Interdisciplinary Social Sciences, Florida State University